

AD-A090 635

AIR FORCE INST OF TECH WRIGHT-PATTERSON AFB OH
AIR FORCE MAINTENANCE TECHNICIAN PERFORMANCE MEASUREMENT, (U)
DEC 79 J R HICKMAN
AFIT-CI-79-241T

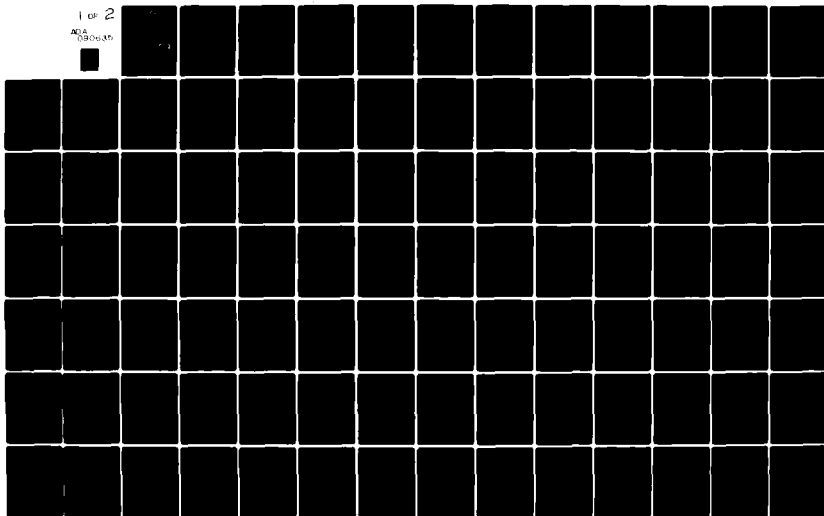
F/6 5/9

UNCLASSIFIED

NL

1 OF 2

AD-A090 635



UNCLASS

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

LEVEL II

①

14 AFIT CI-REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS
BEFORE COMPLETING FORM

1. REPORT NUMBER 79-241T	2. GOVT ACCESSION NO. AD-A090 635	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Air Force Maintenance Technician Performance Measurement.		5. TYPE OF REPORT & PERIOD COVERED THESIS/DISSERTATION
7. AUTHOR Capt Joel R. Hickman		6. PERFORMING ORG. REPORT NUMBER
9. PERFORMING ORGANIZATION NAME AND ADDRESS AFIT STUDENT AT: Arizona State Univ		8. CONTRACT OR GRANT NUMBER(s)
11. CONTROLLING OFFICE NAME AND ADDRESS AFIT/NR WPAFB OH 45433		12. REPORT DATE 28 Dec 79
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) (12) 1-1		13. NUMBER OF PAGES 136
16. DISTRIBUTION STATEMENT (of this Report) APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED		15. SECURITY CLASS. (of this report) UNCLASS
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
18. SUPPLEMENTARY NOTES APPROVED FOR PUBLIC RELEASE: IAW AFR 190-17 25 SEP 1980		DTIC ELECTE OCT 21 1980 S D E
19. KEY WORDS (Continue on reverse side if necessary and identify by block number)		
20. ABSTRACT (Continue on reverse side if necessary and identify by block number) ATTACHED		

AD A090635

3DC FILE COPY

ABSTRACT

Title: AIR FORCE MAINTENANCE TECHNICIAN PERFORMANCE MEASUREMENT

Author: Joel R. Hickman, Capt, USAF

Date: 1979

Number of Pages: 136

Degree Awarded: Master of Science

Institution: Arizona State University

→ The purpose of this study is to find or develop some method for evaluating and measuring the performance of aircraft maintenance technicians in the United States Air Force. This evaluation method is to be used in another research effort to develop a model or models for predicting or evaluating the effectiveness of maintenance technician performance.

The performance appraisal method developed in this study is based on a review of the literature on the subject. A literature review has been necessary, as existing appraisal methods either are not applicable to statistical analysis, are highly inflated, or provide incomplete and non-current coverage of maintenance organizations. The performance appraisal method developed relies on subjective supervisor appraisals of maintenance technician quantity and quality of performance.

An evaluation of the performance appraisal method has been conducted within the aircraft maintenance organization →

→ of one pilot training base. The random sample consists of 20% of the assigned technicians. Thirty-six supervisory groups of five or fewer technicians per group have been selected and found to represent the organization as a whole in terms of experience and relative manning. Quality of performance ratings have a mean value of 7.2 (median of 8.0) on a 10.0 scale, while quantity of performance ratings have a mean value of 6.6 (median of 7.0).

The quality of performance data shows only marginal correlation with existing personnel inspection data. The performance ratings as a whole, however, display superior face validity and usefulness compared to existing personnel inspection data.

Accession For	
NTIS GRA&I	<input checked="checked" type="checkbox"/>
DDC TAB	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By _____	
Distribution _____	
Classification Codes	
Dist	Accession and/or special
A	

AIR FORCE MAINTENANCE TECHNICIAN
PERFORMANCE MEASUREMENT

by

Joel R. Hickman
has been approved
December 1979

APPROVED:

Wenitt H. Young Chair
David D. Bednarek
Lawrence C. Kellie

Supervisory Committee

ACCEPTED:

David D. Bednarek
Department Chair

C. R. Hickman
Dean, College of Engineering
and Applied Sciences

ABSTRACT

Joel R. Hickman, Arizona State University, December, 1979. Air Force Maintenance Technician Performance Measurement. Major Professor: Hewitt H. Young, Ph.D.

The purpose of this study was to find or develop some method for evaluating and measuring the performance of aircraft maintenance technicians in the United States Air Force. This evaluation method was then to be used in another research effort to develop a model or models for predicting or evaluating the effectiveness of maintenance technician performance.

The performance appraisal method developed in this study was based on a review of the literature on the subject. A literature review was necessary, as existing appraisal methods either were not applicable to statistical analysis, were highly inflated, or provided incomplete and non-current coverage of maintenance organizations. The performance appraisal method developed relied on subjective supervisor appraisals of maintenance technician quantity and quality of performance.

An evaluation of the performance appraisal method was conducted within the aircraft maintenance organization of one pilot training base. The random sample for the evaluation consisted of 20% of the assigned technicians. Thirty-six supervisory groups of five or fewer technicians per

group were selected and found to represent the organization as a whole in terms of experience and relative branch manning. The resultant quality of performance ratings had a mean value of 7.2 (median of 8.0) on a 10.0 scale, while quantity of performance ratings had a mean value of 6.6 (median of 7.0). These skewed results presented potential difficulties for regression modeling and for the comparison of distributions. However, these difficulties were overcome for regression modeling, while the quantity and quality distributions were found to be significantly different.

The quality of performance data showed only marginal correlation with existing personnel inspection data. In addition, the use of numbered gradations on the performance appraisal scales resulted in performance histograms which were not useable in most non-parametric tests and which reduced the power of parametric tests for comparisons. The performance ratings as a whole, however, displayed superior face validity and usefulness compared to existing personnel inspection data.

PREFACE

Although this study was conducted with official permission, the findings and conclusions expressed are those of the author, and are not to be construed as official or reflecting those of the United States Air Force.

Many persons have assisted me greatly throughout this undertaking and those mentioned specifically are by no means the only persons who have been helpful. I wish to express my deepest appreciation to the following people:

To Dr. Hewitt H. Young, Professor and Chairperson of my research committee, whose generous support and expert guidance contributed to the completion of this report.

To Dr. David D. Bedworth and Dr. Dwayne A. Rollier, for their comments and assistance while serving on my research committee.

To Col. Walker and Capt. Jerry Raney and the maintenance personnel of Williams AFB, Arizona, who contributed their time and expert help.

To my wife, Gwenne, who labored as the editor of this manuscript.

TABLE OF CONTENTS

	Page
LIST OF TABLES	x
LIST OF FIGURES	xi
 Chapter	
1. INTRODUCTION	1
Purpose	2
2. THEORETICAL PERFORMANCE MEASURES	5
Organization Structure	5
Quality of Ratings	7
Acceptability	13
Relevance	14
Performance Criteria	15
Objective Measures	16
Subjective Measures	18
The Best Traits	21
Appraisal Methods	22
Comparative Procedures	23
Absolute Standards	24
Management by Objectives	27
Direct Measures	28
Suggested Methods	29
Rating Scale Errors	30
The Error of Leniency	30
Error of Central Tendency	31

Halo Effect	31
Scale Format	33
Rules for Writing Scales	33
Rating Standards	33
Scale Anchors	34
The Form of Rating Scales	35
The Rater	37
Supervisory Appraisal	37
Peer Appraisal	38
Self and Subordinate Appraisal	39
Outside Appraisal	39
Recommendations and Conclusions	41
Conclusions	41
3. METHODOLOGY	45
The Data Gathering Instruments	45
Summary of Data Needs	46
Sample Selection Procedures	47
Data Collection	48
Data Analysis	50
Summary	53
4. ANALYSIS	55
Sample Composition	56
Sample Statistical Properties	58
MSEP Data	58
Normality and Applicability to Regression Analysis	60
Rating Comparisons	73

Rating Associations	81
Opinion Survey Analysis	83
Summary	85
5. DISCUSSION	86
The Performance Appraisal Method	86
Evaluation of the Rating Method	88
Test Sampling Procedures	89
Rating Distributions	90
Rating Distributions and Normality	92
Quantity and Quality	92
Validity and Association	94
Maintenance Officer Opinions	96
Summary	96
6. CONCLUSION	99
Contributions and Future Considerations	102
Final Comment	104
REFERENCES	105
APPENDICES	
A. MAINTENANCE TECHNICIAN SURVEY PRIVACY STATEMENT	110
B. QUALITY OF PERFORMANCE RANKINGS	111
C. QUANTITY OF PERFORMANCE RANKINGS	112
D. DATA	113
E. STATISTICAL DATA	118
F. CHI-SQUARE GOODNESS-OF-FIT CALCULATION	122
G. EFFECT OF SKEWNESS AND KURTOSIS ON TYPE I ERROR	123

H.	F-TEST FOR EQUIVALENCE OF SAMPLE VARIANCES	124
I.	COMPARISON OF SAMPLE DISTRIBUTIONS	125
J.	CORRELATION ANALYSIS	127
K.	MAINTENANCE OFFICER SURVEY	130
L.	SURVEY ANSWERS	132
M.	GLOSSARY	135

TABLES

Table	Page
1. Frequency of Rating Categories	20
2. Amount of Disagreement Among Judges in Estimating the Traits of Others	21
3. Validity Coefficients for Graphic Rating Scales and Forced-Choice Sections for Various Criteria	26
4. Grade Distribution of Sample and Population	63
5. Relative Squadron Branch Strengths versus Relative Sample Branch Strengths	64
6. Ratio of Personnel Awaiting MSEP Action/Total Personnel (Maintenance).	69
7. Chi-Square Normal Distribution Goodness-of-Fit Tests	74
8. Skewness and Kurtosis of Sample Data	75
9. Comparison of Variance Equality	77
10. Comparison of Means	80
11. Summary of Correlation Coefficients (r) for Quantity/Quality Associations and Quality/MSEP Associations	82

FIGURES

Figure	Page
1. Deputy Commander for Maintenance Organization	8
2. Organizational Maintenance Organization	9
3. Field Maintenance Organization	10
4. Avionics Maintenance Organization	11
5. Munitions Maintenance Organization	12
6. Illustrations of Conventional Rating Scaling Formats for a Single Item	25
7. Response Distributions Based on Scale Anchors	32
8. Deputy Commander for Maintenance Organization, Williams AFB, Arizona	57
9. Organizational Maintenance Squadron Organization	58
10. Field Maintenance Squadron Organization	59
11. Propulsion and Fabrication Branches, FMS	60
12. Aerospace Systems Branch, FMS	61
13. AGE and Avionics Branches, FMS	62
14. FMS Histograms	67
15. OMS Histograms	68
16. FMS Quality Ratings and MSEP	71
17. OMS Quality Ratings and MSEP	72

Chapter 1

INTRODUCTION

One of the greatest needs of managers of the military weapons system maintenance complex is to measure accurately how well individuals perform on the job. Individual job performance forms one of the bases for performance by the entire organization. If the effectiveness of weapons system maintenance is to be improved, then individual performance must also be measurable and subject to improvement. As stated by Cumminas and Schwab (1973:56), in general "the measurement and assessment of human performance is crucial to effective utilization in order to provide the basis for feedback into the input-processing and input-conversion stages..." of the organizational control process.

Quantifying job effectiveness is, however, difficult. Campbell et al. (1970:101) feel that "Quantifying job effectiveness has been industrial psychology's major bugaboo since its inception." Decades of research by psychologists and personnel experts have failed to provide definitive answers to the question of how to measure performance or effectiveness. Air Force Manual 66-1 (AFM 66-1), Volume I, Maintenance Management (1975:A3-2), allows that the measures of personnel performance form the basis for capability predictions. These measures are, however, difficult to assess and subject to a number of variables. As a

substitute for personnel performance measures, overall maintenance support to the unit's mission is assessed. Such an approach is justifiable given the thirty thousand tasks reported by Wiley (1978:5) that Air Force maintenance performs. Existing official supervisor ratings (e.g., Airmen Performance Ratings) do not serve the performance measurement purpose either as they are general in nature, are not specifically related to tasks and jobs, are highly inflated, and do not discriminate among individuals.

This study considers the available rating techniques, recommends a particular rating technique, and reports on a test of the recommended technique. Chapter 2 will discuss performance and will conclude with a suggested rating scheme. Test methodology will be provided in Chapter 3, Chapter 4 will report on the analysis of test results, and Chapters 5 and 6 will contain an interpretation and a summary of the test results.

Purpose

The purpose of this study is to find or develop a method for evaluating and measuring the performance of aircraft maintenance technicians in the United States Air Force. This evaluation method will ultimately be used as a performance measure of maintenance manpower effectiveness in a research effort to develop a model or models for predicting and evaluating the effectiveness of maintenance technician performance (see Young;1978:15).

The performance rating used must involve minimum development time and cost. These limitations restrict the approaches that can be used. The primary approach used here is a review of published material dealing with performance, with an emphasis on previous studies of Air Force maintenance activities. The recommended performance rating method will then be tested.

Besides cost and time restrictions, any performance evaluation method should meet the following criteria:

1. Be useful for describing performance to management.
2. Be valid as a measurement of maintenance technician performance.
3. Be applicable to different types of performance tasks, such as repair, service, and preventive maintenance.
4. Be applicable to both military and civilian employees of the Air Force.
5. Provide a performance measure throughout the many levels of weapon systems maintenance.
6. Provide valid information for statistical analysis in the form of normal performance distributions with constant variance.

These objectives impose severe restrictions on any possible measurement system. However, satisfying such restrictions is imperative if any research effort is to provide an accurate analysis of the motivation and ability

factors affecting performance. As Guion (1965:90) has stated, "interest should be focused upon what is to be predicted."

In short, the purpose of this study is to answer the following questions:

1. What is the best research method for evaluating or measuring performance of aircraft maintenance technicians in the United States Air Force?
2. Does this method for evaluating or measuring performance provide useful and valid statistical data?

Chapter 2

THEORETICAL PERFORMANCE MEASURES

This chapter deals with the available methods for evaluating and measuring performance based on a review of the literature. The following considerations will be discussed:

1. Organization structure.
2. Quality of ratings.
3. Performance criteria.
4. Appraisal methods.
5. Rating scale errors.
6. Scale format.
7. The raters.

A suggested rating scheme based on the above considerations will be provided at the conclusion of Chapter 2.

Organization Structure

The Air Force maintenance structure involves thousands of personnel performing a vast variety of functions. Thus, any performance measure must be applicable to different organizational levels. This is a difficult requirement to satisfy, as McDonnell (1979) reports that there are forty-five thousand Air Force members in the aircraft maintenance field alone.

Maintenance is concerned with aircraft and missiles

and is performed by military or civilian technicians of both sexes. The three overall levels of maintenance organization are known as base or organizational, intermediate, and depot. Base level maintenance consists of inspecting, servicing, and replacing parts. Intermediate level maintenance is often indistinguishable from base level maintenance and consists of calibrating or replacing damaged or unserviceable parts, of modifying material, and of emergency manufacturing of unavailable parts. Depot level maintenance augments stocks of serviceable material with more extensive shop facilities and personnel of higher technical skill level (usually civilian employees). Although the present research will include only base level organizations, provision must be included for making the proposed performance measurement technique applicable to all levels for further evaluation.

Further generality of the rating technique is mandated by the varied tasks performed by a base level maintenance organization. A typical Air Force base with a mission involving aircraft might include field maintenance (FMS), organizational maintenance (OMS), avionics maintenance (AMS), and munitions maintenance (MMS) squadrons (see Figures 1, 2, 3, 4, and 5). Meister, Finley, and Thompson (1971), Foley (1974), and Wiley (1973) have considered automatic flight control maintenance performance in the AMS alone, while Sauer, Campbell, and Potter (1977) dealt with Short Range Attack Missile maintenance in the MMS alone. Enlarging the

scope of a performance measurement tool to include repair, fabrication, and preventive maintenance personnel as well as flightline launch and recovery personnel, requires either generalized rating scales applicable to many technician specialties, or specific, noncomparable measures for each specialty. Separate measures would, however, make any analysis of overall performance within a squadron impossible.

The nature of the maintenance organization strongly favors the use of general individual performance measures. Such measures would be applicable to the varied tasks and functions for which the different technicians are responsible. Since most maintenance is performed by teams of five to ten technicians working under one supervisor, the supervisor could evaluate his personnel if a general, subjective performance measure were to be used. Thus due to the structure, size, and complexity of the Air Force maintenance system, the present research effort must use a new, subjective, and generalized performance measurement system.

Quality of Ratings

A performance measure is successful, according to Barrett (1966:12), only if it meets three standards:

It must be acceptable to the people who use it; it must cover what is important and only what is important; and a systematic examination of the results of ratings must show that they are reasonably free from important defects.

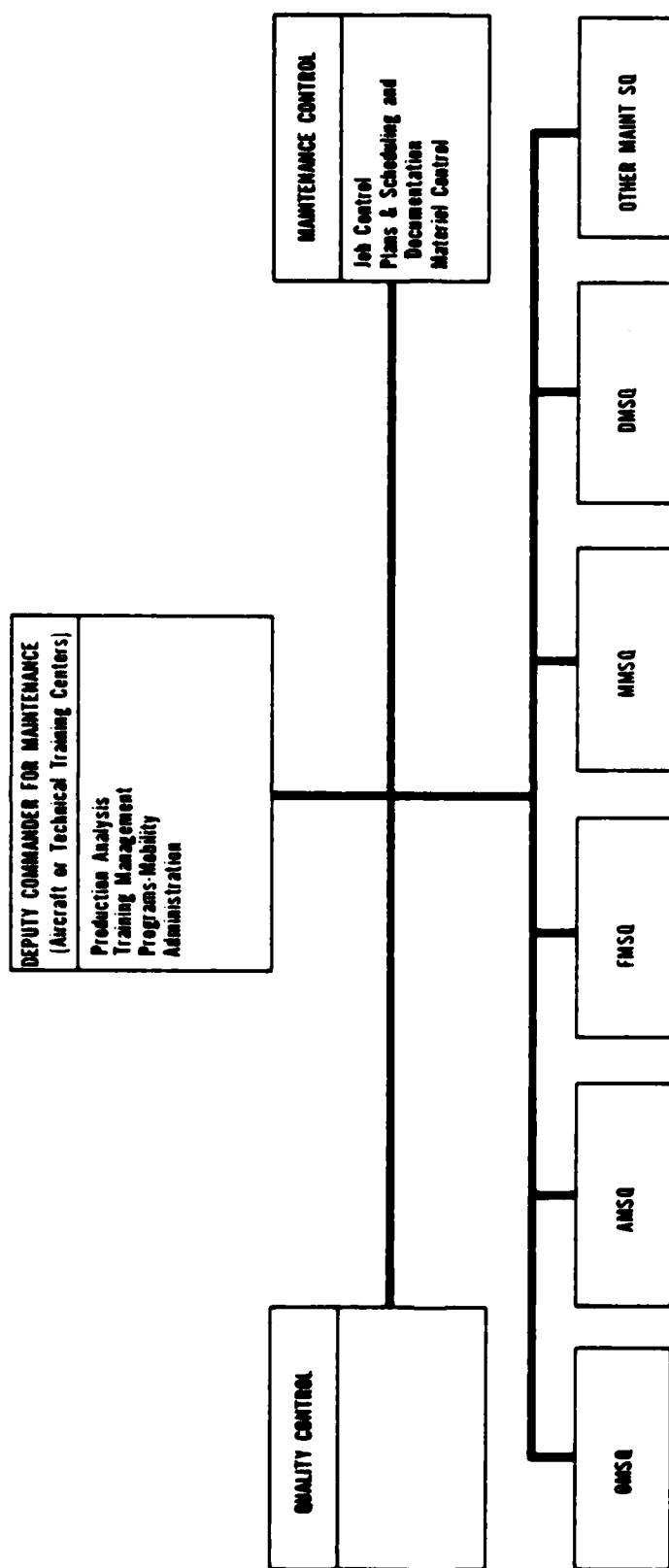


FIGURE 1
Deputy Commander for Intelligence Organization
Source: AFP 6-1, Vol. 1, November 1, 1970.

Source: APD 6-1, Vol. 1, November 1, 1975.

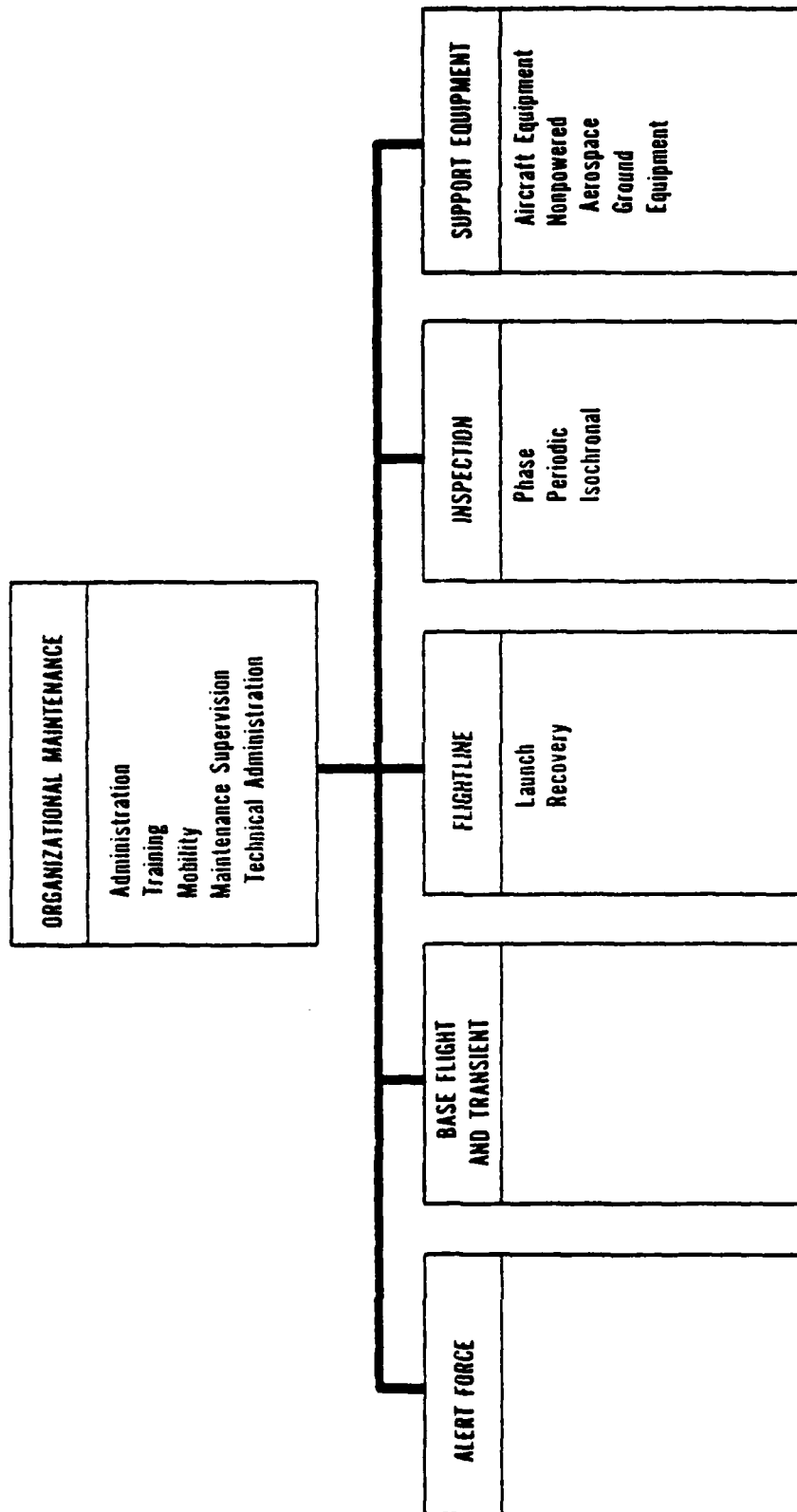


Figure 2

Organizational Maintenance Organization

Source: AFM 66-1, Vol. 1, November 1, 1975.

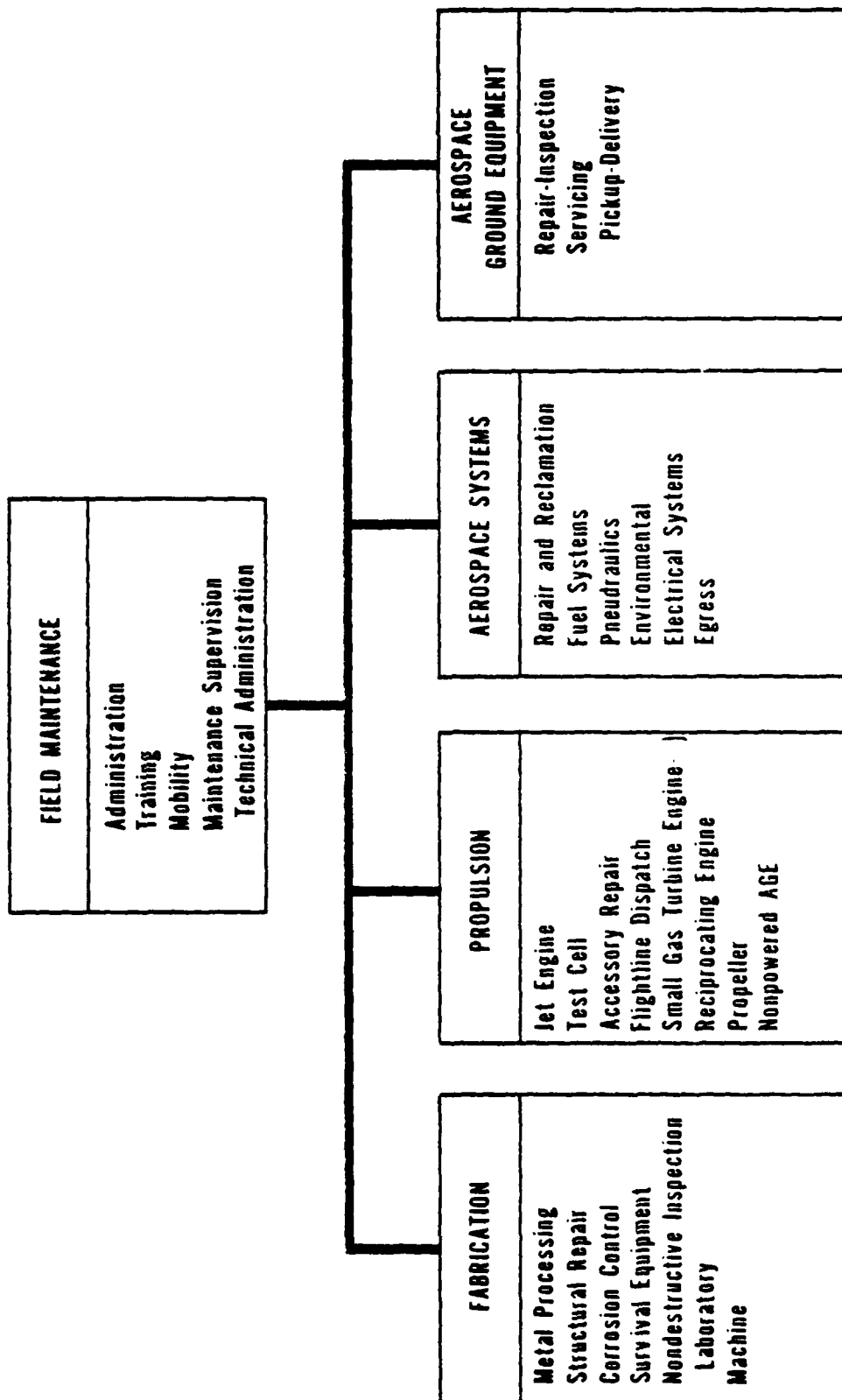


Figure 3

Field Maintenance Organization

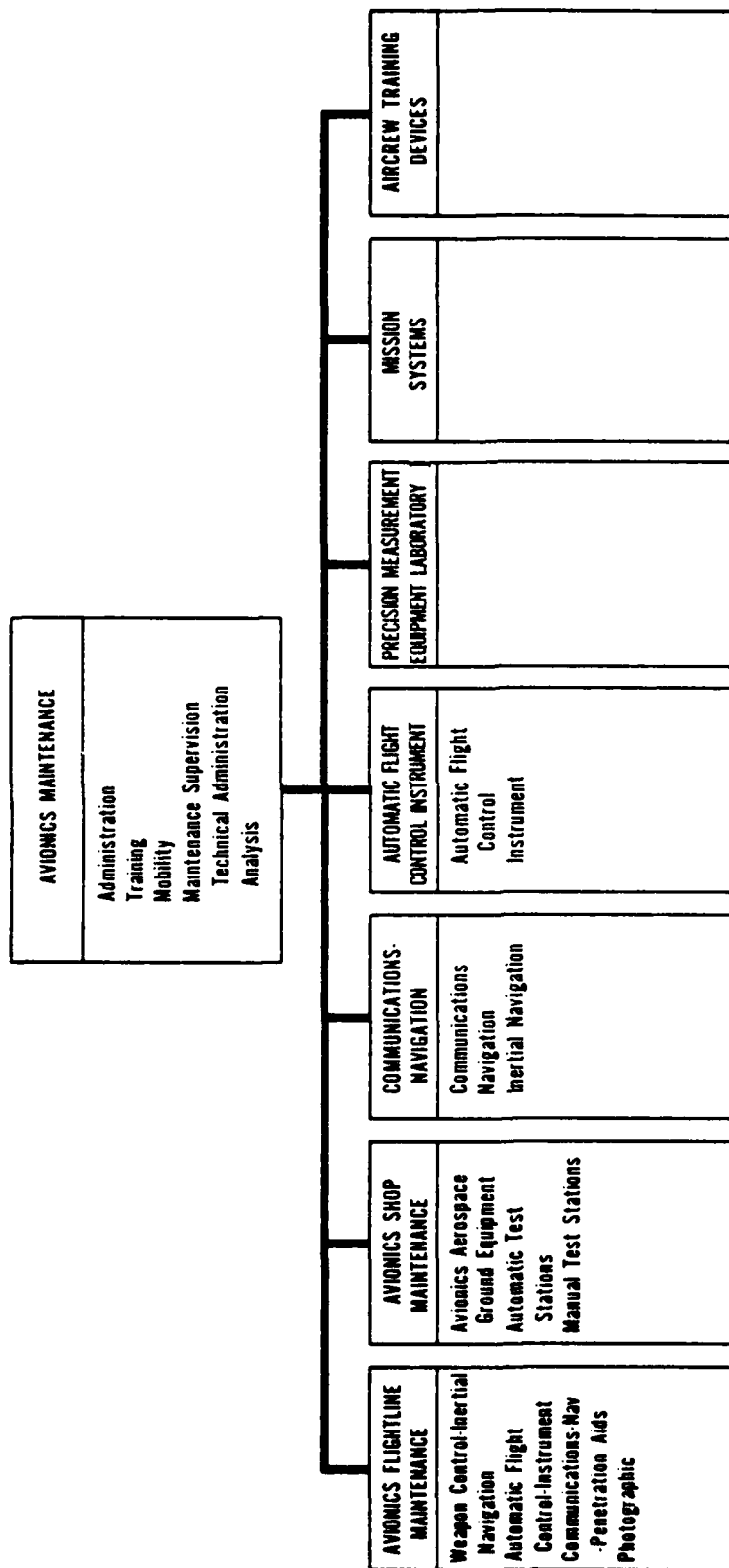


Figure 4

Avionics Maintenance Organization

Source: AFM 66-1, Vol. 1, November 1, 1975.

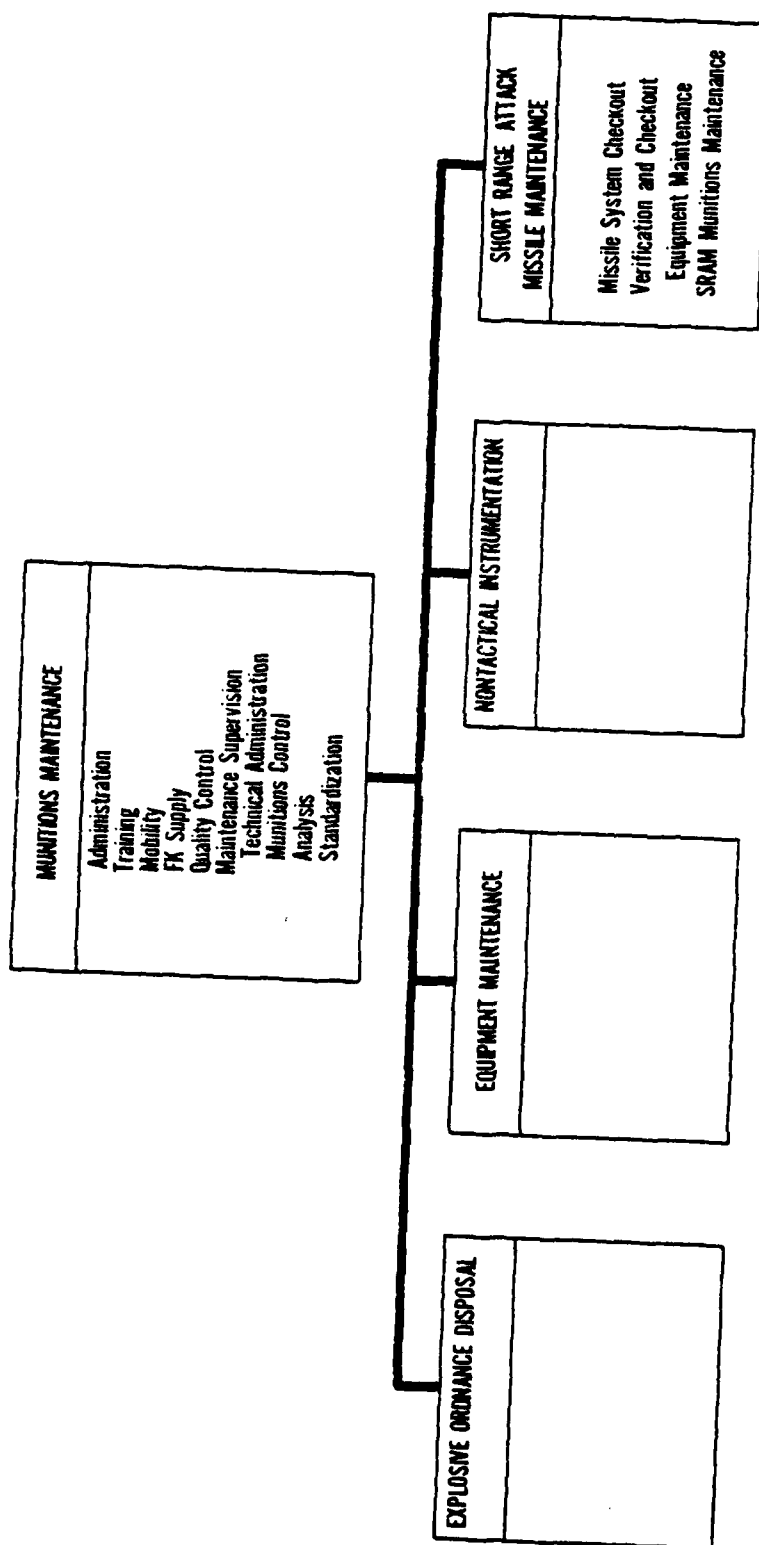


Figure 5

Munitions Maintenance Organization

Source: AFM 60-1, Vol. 1, November 1, 1975.

Acceptability

The performance data which will eventually be used to develop performance effectiveness models must be accepted by maintenance managers and evaluators as well as research personnel. The easiest way to gain acceptance might be to use existing measures such as Airmen Performance Ratings (APRs) or Merit Ratings for civilian personnel. However, these measures are used for the administrative purposes of promotion and wage administration, and not for developmental purposes. McGregor (1957) and Barrett (1966) warn against mixing such incompatible purposes in one program, as management is placed in the incompatible role of judge and counselor.

If a new performance measure is to be developed, it might be advisable to solicit the opinions of managers by using surveys or limited acceptance tests as to criterion utility. An alternative to either using existing measures or soliciting manager opinions as to acceptability would be to develop criterion-referenced test measures. A criterion-referenced test measures what an individual can do, or knows, compared to what he must be able to do, or must know, in order to complete a task successfully (Glaser and Nitko, 1971; Swezey and Pearlstein, 1975). Such Criterion-Referenced Job Task Performance Tests (CJTPT) were experimentally developed by Foley (1974) for electronic maintenance tasks after much time and effort. Such objective tests might prove to be more acceptable than subjective performance

judgments, such as supervisors' ratings.

Relevance

Acceptance is not enough; a measure that omits essentials or gives weight to trivia is defective. Barrett (1966) feels that a clear statement of the objectives of the ratings is the first step, while Guion (1965) believes that the first step is a judgment of the importance of the concept being developed. Both authors agree that the second step is a clear statement of what the job requires and the kinds of job behavior that are essential to success. As Barrett points out, punctuality may be important in an automated office where each person's performance affects his neighbors, but it is unrelated to the success of a door-to-door salesman.

In deciding whether a rating is relevant, it is helpful to check it against standards described by Brogden and Taylor (1950). The three defects they identify are deficiency, contamination, and distortion.

Deficiency. This defect results if the measure of performance lacks any elements necessary to give adequate coverage. Rating or ranking of "overall performance" gives the illusion that everything is included while, in fact, raters may have different concepts of job elements and different ideas of what constitutes successful performance. Cummings and Schwab (1973:46) also consider measurement deficiency to exist if employee productivity is accounted

for by quantity of output alone without also considering quality of output.

Contamination. Lopez (1968:211) feels that contamination occurs when behavioral characteristics that are unrelated to job performance are included in an evaluation method. Such unrelated characteristics include "self-confidence," "self-control," and "personality."

Distortion. When several criteria are used to define performance it is possible to distort their importance by improper weighting. Criteria which are not specific may allow inclusion of dramatic or easily observed events such as frequent tardiness or a lucky break in the evaluation.

All of these defects can be avoided with careful selection of the performance criteria to be evaluated. Procedures for selecting such criteria will be discussed next.

Performance Criteria

The ideal topics for rating must be both important and ratable. As Barrett (1966:33) points out, these two attributes do not necessarily go together, as some trivial areas such as regularity of haircuts may be accurately rated while important concepts such as output and quality are harder to pin down.

In general, Lopez (1968:37) believes that performance refers to a specific kind of human behavior in a "system"

environment and activity. He feels that some employee performance evaluation procedures are designed to judge only behavior, while others are designed to judge only results. The first approach is too general and the second too narrow because the proper object of the process is the evaluation of the act of performing in terms of both results and behavior.

Guion (1965:91-96) indicates that two types of criteria can be used. These are objective measures of job behavior and judgment ratings. Objective or countable measures of behavior can be grouped into two major categories: production data and personnel data. "Production data" includes quantity and quality of output, while "personnel data" includes absence or accident rates.

Objective Measures

Attempts to use objective data in analyzing maintenance performance were made by Sauer, Potter, and Campbell (1977), Foley (1974), and Meister, Finley, and Thompson (1971). Sauer, Campbell, and Potter (1977:22) attempted to use individual task performance for Short Range Attack Missile (SHRAM) technicians through the Strategic Air Command (SAC) Maintenance Standardization and Evaluation Program (MSEP). This provided information on technician performance against standards for technical errors, safety errors, and reliability errors. Technician tasks, however, are designed for ease of completion, which results in

very few errors and limited variability. These performance measures are thus of limited value for computing the relationship between human resource factors and task performance.

Meister, Finley, and Thompson (1971:31), utilized observers to record the performance of technicians on a very specific electronics maintenance task--autopilot repair. Two types of performance variables were recorded: those which were based on objective observation (e.g., elapsed time, error frequency, number of components removed and replaced), and those which were based on the subjective judgment of the observer and the observed technician (e.g., efficiency of performance, difficulty of task). The drawbacks of this method include the need to train observers for particular maintenance functions and the lack of relevance of the measures for service functions (e.g., refueling, canopy cleaning, etc.) performed by Organizational Maintenance Squadron personnel.

Foley (1974) advocates the use of Criterion-Referenced Job Task Performance Tests (JTPT). Ronan (1976) reports that a Task Performance Test for firemen led to the adoption of nine independent performance factors, which are superior to peer and supervisory subjective evaluations. Such systems are difficult, costly, and time-consuming to develop, according to Obradovic (1979). No such rating measures now exist for the many maintenance tasks performed by Air Force technicians.

Subjective Measures

Subjective ratings or judgments are relied upon by management as criteria for validation studies. Guion (1965:96) reports that eighty-one per cent of validation studies appearing in the Journal of Applied Psychology and Personnel Psychology between January, 1950, and July, 1955, relied upon ratings.

According to Barrett (1966:33), rating scales are concerned with three kinds of concepts: personality, performance, and product. Personality is the total of a person's characteristics. It includes emotional make-up, intelligence, and what is commonly called character. Performance has to do with how an individual goes about doing work. Included are working hard, following instructions, planning, and taking responsibility. Product is a person's output. The quantity and quality of work are product.

The most pertinent of the three is product. Management is fundamentally interested in sales, production of finished goods, and other factors that are visible and inherently measurable. Product in some cases can be measured directly (objective measurement) and in other cases it is necessary to have a rater look at the product and evaluate its quality. Measures of product often suffer from deficiency, as only part of an individual's output can be measured in objective terms. They may also be contaminated, since much of what is measured is beyond the individual's control; for example, product may be the output of many

individuals, not one alone.

Existing ratings of individuals employed by the Air Force are of little value except for administrative purposes. Airmen Performance Ratings (APRs) are inflated, according to Callander (1979), and are of little value as a single performance measure. Civilian and military personnel appraisals are also privileged information which are difficult to gain access to.

If production is not available for evaluation, the rater may evaluate how the employee goes about his work, instead of what he produces. Though not as objectively measured as products, these job performance characteristics are both ratable and important. Studies by Barrett (1961) indicate that supervisors and subordinates are quite sensitive to performance, agree on the relative importance of performance traits, and attach a great deal of weight to the performance style used on the job.

Most nebulous, but frequently rated, is personality. Employees are expected to be trustworthy, loyal, helpful, friendly, courteous, kind, and reverent. However, no one knows which of these characteristics contribute--and how much they contribute--to job success. Indeed, agreement on definitions of traits is much harder to reach than agreement on product or performance.

A survey of fifty merit rating plans by Habbe (1956) shows that the element of personality, the most difficult to rate, was the most widely used. The rating of product

(using Barrett's definition) was confined to quantity and quality of output. The findings are summarized in Table 1. Holley, Feild, and Barnett (1976:459) reported similar results on the frequency of category use.

Table 1
Frequency of Rating Categories (Habbe, 1956)

Category	Freq.	Category	Freq.
Group 1: The Old Standbys (Product)			
Quantity of work	44	Quality of work	31
Group 2: Job Knowledge and Performance			
Knowledge of job	25	Safety habits	7
Attendance	14	Good housekeeping	3
Punctuality	12		
Group 3: Characteristics of the Individual (Personality)			
Cooperativeness	36	Initiative	27
Dependability	35	Intelligence	17

The major emphasis of ratings should be on the product of an individual's effort in terms of what he or she accomplishes. When there are no products, performance is suggested as being the next best level of abstraction to deal with, while pure personality variables have little if any relevance to the performance measurement task. Hollingworth (1922:79) provides evidence that some traits

are more reliably measured than others. Only personality traits were studied and Table 2 summarizes the relative disagreement between judges concerning traits.

Table 2
Amount of Disagreement Among Judges in Estimating
the Traits of Others (Hollingworth, 1922)

Trait	Divergence	Trait	Divergence
Close Agreement			
Efficiency	83	Perseverance	85
Originality	86	Quickness	89
Fair Agreement			
Breadth	96	Intensity	99
Leadership	96	Reasonableness	100
Poor Agreement			
Courage	109	Integrity	117
Unselfishness	110	Cooperativeness	119

The Best Traits

In this case it appears that subjective appraisals are most applicable. There are, however, many potential traits that could be used. Lawler (1967:371) indicates that it is easy to err on the side of providing too many traits upon which to make ratings. Dunnette (1963:252) points out that the use of a single criterion is unrealistic, while

Rush (1953:23) indicates that between three and five criterion factors surface in factor-analysis studies. The potential size of a study covering Air Force maintenance performance mandates the use of as few factors as possible.

Lawler (1967:371) indicates that one rating that probably should be included is one on quality of job performance. When people are asked to make such general ratings on quality they act in a very predictable way, as efficient appraisers of critical incident data from their observations of an individual's performance in the past. The other traits besides quality that should be used in performance analysis are difficult to specify. They should be based on the purpose of the study and on particular types of behavior that characterize the important functions of the job. Wiley (1978:23) included quantity of work, self-initiation, sharing of knowledge, and exceeding one's share as additional rating dimensions. In this study, quantity and quality of output are applicable to all technician functions and are of interest to management.

Appraisal Methods

A wide variety of appraisal methods has been developed. The major appraisal methods come under four general headings: (1) comparative procedures, (2) absolute standards, (3) management by objectives (MBO), and (4) direct indexes.

Comparative Procedures

Comparative procedures are frequently characterized by two features. First, the evaluation is made by comparing one individual against another on the particular dimension of interest. Second, this comparison is often made on a general dimension which attempts to measure an employee's overall contribution to the organization. Two popular comparative procedures are straight ranking and paired comparison.

Straight Ranking. In an appraisal context the evaluator is typically asked to consider all of the employees to be appraised and identify the very best performer, the second best, and so on through all employees to the very poorest. Cummings and Schwab (1973:82) feel that this procedure is natural for most evaluators, as people are frequently informally ranked. Barrett (1966:46) indicates that ranking is free of leniency and central tendency but the ability to show relative performance between people is lost. Sauer, Campbell, and Potter (1977) used a ranking procedure with a conversion to normalized percentiles as described by Guion (1954:181) to analyze maintenance personnel performance. This procedure is based on the assumption that performance is normally distributed over a population sample.

Paired Comparisons. This system requires the evaluator to compare each employee to be ranked with every other

employee, one at a time. An employee's standing in the final ranking is determined by the number of times he or she is chosen over the other employees. This system can be tedious and result in a large number of comparisons.

Absolute Standards

With appraisal systems using absolute standards, individuals are evaluated against one or several written standards. There are two general absolute standards methods. First, qualitative methods, where the evaluator is asked to identify whether the appraisee possesses or does not possess, in a qualitative sense, some performance characteristic. And secondly, quantitative methods, where the evaluator attempts to measure the degree to which each appraisee possesses certain characteristics.

Qualitative Methods. Critical incidents and forced choice are illustrative of qualitative methods. Flanagan (1949) describes the critical incident method as a method that provides a picture of individual performance. The rater records on a special form examples of outstandingly good and poor performance on the part of the individual. This method is not useable in this study as it would provide nebulous results and be cumbersome to evaluate with many maintenance technicians.

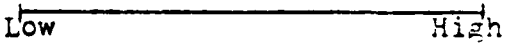
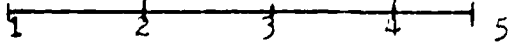
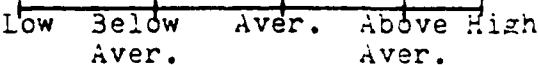
Forced choice procedures involve a series of groups or clusters of statements about job behavior. The evaluator is asked to choose the item which is most descriptive of the

appraisee. Travers (1951) notes that forced choice makes it difficult or impossible for a person to control the quality of the rating. The descriptive statements of job behavior must be developed for each individual job, a procedure that is also not useable in this research situation.

Quantitative Methods. Conventional rating procedures and behaviorally anchored rating procedures are examples of quantitative methods. According to Locker and Teel (1977:246), conventional ratings constitute the most popular form of appraisal techniques. Rating scales generally have several statements about employee characteristics or behavior. A continuous or discrete scale is established for each item. Figure 6 illustrates several scaling procedures from Cummings and Schwab (1973:90). Item A is scaled continuously:

Figure 6

Illustrations of Conventional Rating Scaling Formats
for a Single Item (Cummings and Schwab, 1973)

Item	Scaling Format
A Overall job performance	
B Overall job performance	
C Overall job performance	

the evaluator places a check somewhere on the scale to represent his assessment of the appraisee. Item B has a numerical discrete scale although letters are sometimes used

instead of numbers. Item C is also scaled discreetly with adjectives. Discrete scales generally result in greater agreement amongst raters and hence are preferable to continuous scales, according to Cummings and Schwab (1973). However, the overall validity of rating scales has been questioned. Bayroff, Haggerty, and Rundquist (1954:105) concluded as a result of some extensive work on Army ratings that "Ratings using different types of rating techniques were not markedly different in validity." Their comparison of graphic scales, forced choice, and a controlled checklist with three criteria is shown in Table 3. It is significant

Table 3

Validity Coefficients for Graphic Rating Scales and
Forced-Choice Sections for Various Criteria
(Haggerty and Rundquist, 1954)

Ratings	Rank by Associates	Class Standing	Efficiency Reports
Graphic Scale: overall value	.53	.35	.19
Graphic Scale: competence for duty assignment	.43	.25	.10
Forced-choice pairs	.41	.25	.10
Controlled checklist	.44	.31	.26

to note that in Table 3 ranking by associates is a superior criterion when compared with the validity of existing performance measures such as class standing or efficiency report scores. Furthermore, overall value graphic scales

are superior to any other rating method investigated.

An alternative quantitative rating method is the use of behaviorally-anchored rating scales (BARS). Millard, Luthans, and Otteman (1976) feel that BARS may represent a substantial improvement over traditional rating approaches. Three basic steps are involved in BARS: (1) critical incidents are used to determine job-related behaviors and important performance dimensions, (2) the job-related behaviors identified in the critical incidents are linked with the appropriate performance dimension, and (3) significant behavioral incidents are numerically scaled to a level of performance. BARS overcome two methodological problems found in conventional ratings: BARS identify the critical item included in an assessment and scale these critical items against specified levels of performance. BARS are not, however, applicable in this study as they require separate scales for individual job responsibilities.

Management by Objectives

Management by Objectives (MBO) has been offered by McGregor (1960) and others as an alternative to conventional rating and employee comparison systems. Wikstrom (1968:2) feels that MBO is based on two related concepts: "(1) the clearer the idea one has of what it is one is trying to accomplish, the greater the chances of accomplishing it; and (2) progress can only be measured in terms of what one is trying to make progress toward." MBO is primarily a

developmental procedure for individuals rather than an evaluative one. As such, MBO is not applicable to this case.

Direct Measures

All of the procedures described to this point require that employee performance be evaluated or assessed by someone. It is also sometimes possible to obtain information about performance more directly without the necessity of the performance behavior being filtered through the evaluative processes of an appraiser.

For instance, it is sometimes possible to measure the productivity of an individual directly. These measures are generally aimed at the quantity (e.g., hourly units of output, monthly gross sales) or quality (e.g., percent units rejected, scrappage) of output. Unfortunately, no universal quality or quantity measures exist for Air Force maintenance. While quantity measures could be developed using industrial engineering job standards, AFM 66-1, vol. 1, (1975:1-7) mandates that standards be developed to evaluate mechanics' performance in only certain recurring tasks. These certain recurring tasks are those which (1) consume a large number of man-hours, (2) involve extremely high cost components, or (3) require a large amount of equipment or downtime. This limited use of standards thus makes quantity direct measures impossible.

Quality control in the Air Force is measured in a

subjective manner, since many maintenance tasks such as refueling or preventative maintenance result in no product subject to a rejection or scrap rate. Furthermore, the personnel evaluations required by AFM 66-1, vol. 10, (1977:4-11) are not completed for each individual on any regular basis. In addition, no sampling procedures are specified to ensure that a representative sample of the technician population is evaluated. As Sauer, Campbell, and Potter (1977) discovered, even the results of Air Force evaluations are not useable in a statistical analysis of performance due to the performance scoring methods used and the resultant high level of performance.

Direct measures, while the least questionable source of performance information, are simply not available as a useable source for statistical analysis. Indeed, the existing quality control system makes it difficult to ensure that a sample representative of maintenance technician performance can be obtained.

Suggested Methods

Of the appraisal methods reviewed, the only applicable methods are straight ranking (a comparative procedure) and rating scales (a quantitative, absolute standard). Both methods are based on subjective appraisals of perceived performance. The use of either method in appraising performance is open to discussion. Evidence indicates that the use of two rating procedures in conjunction with each other

increases the accuracy of the final rating since the rater is forced to carefully consider each appraisee for the first rating procedure before giving his final rating. Campbell, Prien, and Brailey (1960:440) concluded that "Graphic scales following a [performance] checklist show higher apparent validities than the [performance] checklist [alone]."

Similar results have been found for graphic scales following a forced choice report, according to Barrett (1966:71). It is suggested that in this study graphic scales follow a forced straight ranking appraisal. This should be applicable for research and provide rating scale performance values which are normally distributed and acceptable for statistical analysis.

Rating Scale Errors

The use of ratings rests on the assumption that the human observer is a good instrument of quantitative observation, i.e., that the observer is capable of some degree of precision and some degree of objectivity. Several observable errors do arise in rating scale use, however. These errors include the error of leniency, the error of central tendency, and the halo effect.

The Error of Leniency

Often ratings tend to cluster about a point at the favorable end of any scale used to appraise personnel. This is due to leniency on the part of appraisers. Barrett

(1966:23) observes that often, when the descriptive word "average" is included on a scale, more than half the appraisees are given ratings above average. This is a logical impossibility if these individuals are truly compared with others in the organization. In order to reduce errors of leniency, Guilford (1954:278) suggests eliminating the word "average" from any scale. According to Bittner (1946), additional ways to reduce errors of leniency include the use of ranking and the review of ratings by several levels of supervisors. No published work could be found concerning the effect of peer appraisals on leniency.

Error of Central Tendency

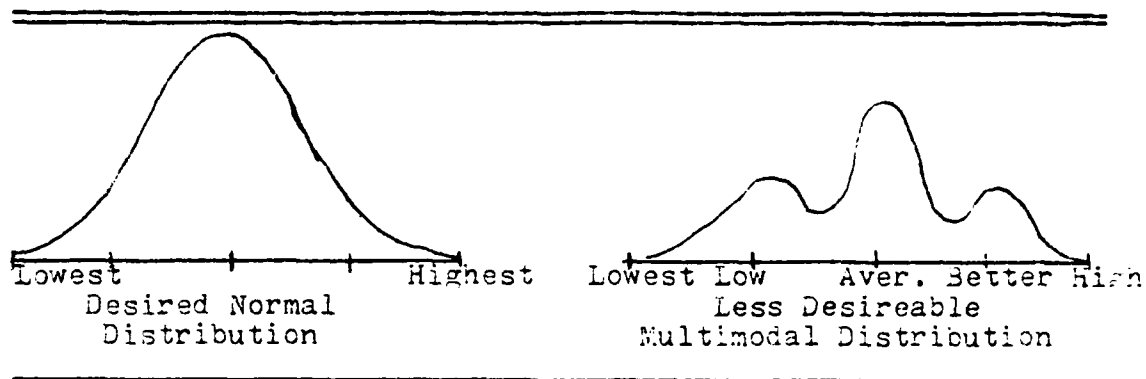
As defined by Guion (1965:90), this error is marked by restricted variability around the center of the scale. Raters tend to put their ratings in the center of the scale when they are not entirely clear as to the meaning of ratings or when they do not know the person they are rating. Clear definition of rating criteria and the use of immediate supervisors reduces this problem. The use of a few descriptive adjectives in the middle of the scale also creates problems, as appraisal distributions tend to be multimodal and non-normal (see Figure 7). No published work could be found which determined if central tendency merely reflects a normal distribution of appraisals over a scale.

Halo Effect

As defined by Guion (1965:90), halo is the tendency

Figure 7

Response Distributions Based on Scale Anchors



to rate an individual in the same manner on all traits because of a general, overall impression that can be either favorable or unfavorable. Halo thus results in positive correlation between the traits that are rated. Halo may be reduced by using a format proposed by Stevens and Wonderlic (1934) that calls for rating all appraisees on one trait, then rating them on the next trait, and so on. Guilford (1954:272) also indicates that one trait per page should be used. Ranking methods, of course, eliminate the halo effect.

In general, the above errors can be avoided by using clear definitions of traits, by concentrating on a single trait at a time, and by avoiding limited descriptive adjectives and words such as "average." It is not known if peer review deflates ratings (i.e., reduces leniency). It is also not known if central tendency errors simply reflect normal distributions of appraisal ratings.

Scale Format

Once it is decided what should be rated and by what means, then this information has to be communicated to the rater or raters so that they know what to do. This problem of communication is critical to the success of any rating scheme. All raters should rate using the same criteria for the same purpose to produce useable results that reflect the performance of individuals in the organization. Although this ideal can never be met when subjective ratings are used, several considerations related to scale building can improve ratings. Among these considerations are rules for writing scales, rating standards, scale anchors, and the form of the rating scales. All of these will be considered.

Rules for Writing Scales

Several authors have provided rules for writing scales. Uhrbrock (1961) provides a useful list of two thousand scaled items. Some of the most important precepts are as follows:

1. Express one, and only one, thought in a scale.
2. Use words the rater understands.
3. Have the raters rate what they observe, not what they infer.
4. Eliminate double negatives.
5. Express thoughts simply and clearly.
6. Keep statements internally consistent.
7. Avoid universal terms such as all, always, and never.
8. Stick to the present.
9. Avoid vague concepts.

Rating Standards

The rater who has been informed adequately of the area

he is to rate is still not equipped to do an evaluation; he must know the standards against which the rating is to be made. These standards are based on the previously discussed consideration of the types of ratings, the purpose of the ratings, and the organizational setting in which they are arrived at. To aid in providing a framework for clearly expressing one thought in a scale, Barrett (1966:77) provides three related standards against which performance is generally rated: comparison with others, comparison with job standards, and comparison with absolute standards. Job and absolute standards do not exist for all tasks performed by Air Force maintenance technicians. As a consequence, the only available standard is comparison with others. In this case the appraisee is evaluated in relation to other people in some specified group. Although any group may be specified, the most pertinent is made up of workers on the same job or on a similar one. Such comparison is made most directly when the rater is asked to rank a group of employees.

Scale Anchors

Scale anchors are numbers, words, or phrases used to tell the rater the significance of making his rating at a given point. Taylor et al. (1958) found that formats incorporating behavioral descriptions of scale steps were superior in reliability to numerically anchored scales. However, a criticism commonly leveled at the use of behavioral descriptions such as "excellent," "highly favorable," "fair," and "poor" is that the words do not have a common meaning.

Careful work by Jones and Thurstone (1955) contradicted this criticism and supported Taylor's findings. They also discovered that scales in which the end points are wider apart give more reliable results than do those in which the spread is constricted. When only end anchors are given, there is less error at the extremes than at the central value.

For the present study, a quality scale could be anchored with the adjectives "lowest" and "highest," while quantity could use "slowest" and "fastest" as anchors. No intermediate scale descriptions are available for the wide variety of tasks performed by Air Force maintenance personnel. Avoiding average scale values or any other adjectives would avoid the problems of multimodal performance distributions, as shown in Figure 7 on page 32.

The Form of Rating Scales

Considerable attention has been paid in experimental psychology to the problems of scaling to find out all that can be learned about man as a measuring instrument. Experience has shown that certain rules are favorable to effective graphic ratings. Guilford (1954:267) lists the following rules:

1. Each trait should occupy a page by itself.
2. The line should be at least five inches long, but not much longer.
3. The line should have no breaks or divisions.
4. The "good" or "high" ends of the lines should be in the same direction.
5. For unsophisticated raters, the "good" end should be placed first.
6. Descriptive phrases or cues should be concentrated as much as possible at points.
7. End cues should not be so extreme in meaning that

- they will never be applied.
8. End cues should be set at a little distance from the ends of the line.
 9. In scoring, a stencil should be used that divides each line into sections to which numerical values are assigned.

The number of steps in a scale varies. Bendix (1954) reports on experiments in ratings in which he found that satisfactorily high reliabilities were obtained on scales involving three to nine levels. The Air Force currently uses ten levels for Airmen Performance Ratings, a form that all military technicians are familiar with. The use of up to ten levels in this case is supported by Barrett (1966:87), who feels that raters can make finer distinctions when the scale calls for judging the differences between two people than in rating a person against a standard.

Barrett (1966:89) also feels that much discussion but little research has centered on the problem of an odd or even number of scale steps. The even-numbered scales deny the rater the use of the term "average" as a rating, the easiest rating to make. Odd-numbered scales, on the other hand, allow average ratings, as there should be more average people than any other kind. There is no conclusive evidence with which to resolve the issue; the presence or absence of a central point when more than five levels are used probably does not make much difference.

It thus appears that the best scale format for this study should follow the rules listed by Uhrbrock and Guilford. The rating standards should be based on comparisons with other technicians within a particular maintenance

squadron. Using two adjectives to anchor the ends of scales for quality and quantity of performance appraisal should serve several purposes: (1) the term "average" would be avoided, (2) generality of the scale would be maintained to make it applicable to many maintenance activities, (3) the possibility of obtaining a normal performance distribution would be improved, and (4) multimodal distributions grouped around descriptive adjectives would be avoided. The use of ten steps should be familiar to the raters due to the similarity with Airmen Performance Ratings and, since interpersonal performance is being rated, should allow for finer distinctions between technicians.

The Rater

There are five possible parties that can do the appraising: (1) the supervisor(s) of the person to be appraised, (2) organizational peers of the appraisee, (3) the appraisee himself, (4) subordinates of the appraisee, and (5) persons outside the immediate work environment of the appraisee. Any of these parties might be appropriate, depending on the purpose (either evaluative or developmental) of the appraisal and the dimensions (either outcomes or methods) being appraised. This study is primarily concerned with evaluative purpose based on outcomes.

Supervisory Appraisal

There are two primary justifications for centering the appraisal process on the performer's superior. The hierarchy

of formal authority which exists in most organizations legitimizes the right of the superior to make evaluative and developmental decisions concerning his subordinates. Lerner (1968) and Vanzelst and Kerr (1953) have shown that the supervisor is the person most employees want and probably expect to appraise them. Thornton (1968) has shown that supervisor ratings are valid, while Barrett (1966) has shown them to be reliable if care is taken to train appraisers and to use an acceptable appraisal form. If supervisors of Air Force maintenance technicians are used as raters, however, a training program or personal supervision of the supervisors as raters would be necessary. While a training program would be difficult and extensive, personal supervision of the raters by the researchers, either in a group or on an individual basis, would serve to enhance the quality of ratings and to make training unnecessary, according to Taylor and Lounsbury (1961). Whitla, Dean, and Pirrell (1953) indicate that those raters on the supervisory level functionally closest to the ratees are best able to rate them. Supervisor ratings are thus the most applicable rating scheme available for rating Air Force maintenance personnel performance.

Peer Appraisal

According to Lewin and Zwary (1976), peer ratings have been empirically shown to have high validity in the prediction of diverse future performance criteria. However, current--not future--performance is of interest in this study. Furthermore, peer appraisals have seldom been used for evaluation

outside of military academies and officer training schools. It thus appears that the use of peer appraisal was little to offer over supervisor ratings of technician performance. It is possible that peer appraisals could be useful in reducing rater leniency, although no reports on research into this subject could be found.

Self and Subordinate Appraisal

Subordinate appraisal of maintenance technicians is not applicable to this study, as technicians do not rate their superordinates. Self-appraisals are also not applicable, as they lack inherent validity and are seldom used for maintenance managers.

Outside Appraisal

Outside appraisal was used by Nelson, et al., and Thomson (1971) as a means to evaluate aircraft performance on aircraft auto-pilot systems. This method is time-consuming and costly, as observers usually need to be trained in appraising a variety of maintenance tasks. In the present study, the number of different tasks involved in the training of observers. Outside appraisers, according to Carrhell et al. (1962), are less lenient than supervisors. This same effect might be achieved through self-appraisal, although no literature appears to exist to support such a theory.

In general it is apparent that the use of immediate supervisors as performance raters is most appropriate for

42

this study. Immediate supervisors are knowledgeable about their personnel and the desired performance in a particular maintenance specialty.

Recommendations and Conclusions

One of the greatest needs of managers of the military weapons system maintenance complex is to measure accurately how well individuals perform on the job. Individual job performance forms one of the bases for performance by the entire maintenance organization. And if the effectiveness of maintenance is to be improved, then individual performance must also be measurable and subject to improvement. Measuring individual performance can be difficult, however.

Conclusions

The greatest problem in measuring individual performance is that existing rating schemes are either not applicable to statistical analysis or are highly inflated and unusable for research. Airmen Performance Ratings and Merit Ratings are used for administrative purposes of promotion and demotion and seldom reflect job performance alone. Proficiency ratings are based either on paper and pencil theory tests which do not reflect job performance or on infrequent observations (NSEP) which are highly skewed and unusable for statistical analysis and which reflect compliance, not performance capability. Since existing data is not applicable, a new rating scheme must be developed.

Any new performance rating scheme must: (1) be valid-

cable to all levels of the maintenance organization, (2) have high quality, useful criteria measures, (3) have valid measures, (4) be free of error, (5) have an accurate format, and (6) be used by an appropriate rater. In the first instance, the size of the Air Force maintenance organization requires a measurement scheme applicable to civilian and military technicians of all races and sexes performing many tasks ranging from servicing aircraft to repairing missile guidance systems. Most maintenance of any type is performed by teams of five to ten technicians supervised by one individual. It thus appears that the organization structure and size restrict performance measures to general criteria such as quantity and quality of performance based on subjective appraisals by immediate supervisors.

Secondly, any new performance measure must be of high quality, i.e., it must have face validity. This means that it must be acceptable to the people who use it, as well as being relevant for management. Acceptability can be determined by using a measure similar to existing measures or by surveying maintenance management personnel concerning performance by technicians. Relevance can be achieved by avoiding deficiency, contamination, and distortion; avoidance of these problems depends, to some extent, on what is rated, how it is rated, and by whom.

Thirdly, the performance rater must have valid measures. Objective rating criteria are simply not available or not useful to this study. Therefore subjective traits or

criteria are the only ones being considered. Of such subjective traits, performance or product criteria appear to be the most valid according to existing literature on the subject. And of those categories available, individual quality and quantity of performance by technicians are the most applicable. These concepts are easy to compare in personnel, provide better agreement between raters, and provide information useful to management. In particular, quantity and quality of performance are easy to relate to individual motivation and capability in an overall performance model.

Fourth, the performance rating must be as free of error as possible. Of the appraisal methods reviewed, the only applicable methods for this study are straight ranking (a comparative procedure) and rating scales (a quantitative, absolute standard). Both methods are based on subjective appraisal of perceived performance. Evidence indicates that the use of graphic scales following a forced ranking increases the accuracy of the ratings. This method should be applicable for research in the present study and should provide rating scale performance values which are normally distributed and acceptable for statistical analysis.

Fifth, the rating scale must have an accurate format. The actual rating scale proposed by this study is designed to minimize errors of leniency, of central tendency, and of the halo effect. The quality and quantity of performance appraisal forms proposed (Appendices B and C) are adapted from Sauer, Campbell, and Potter (1977). The directions have

been edited to conform to Uhrbrock's (1961) rules for appraisal forms, i.e., thoughts are expressed clearly and simply, statements are internally consistent, and words that the raters understand are used. To conform with Guilford's rules (1954), each trait occupies a page by itself, the scale line is five inches long, and ten steps corresponding to increasing performance compared to other technicians in a squadron are used to provide an adequate spread of responses.

Finally, the performance rating must be used by an appropriate rater. For this study the use of immediate supervisors as performance raters is the most appropriate technique. Personal supervision of supervisors, either in a group or on an individual basis, serves to enhance the quality of ratings and to make training unnecessary. It might also be applicable to have peers complete appraisals in an attempt to reduce rater leniency.

In short, the suggested rating forms in Appendices B and C are the best that can be developed based on a review of the literature concerning appraisals and on the nature of the Air Force maintenance organization. These rating forms have face validity, if previous research conclusions are accepted. It remains to be seen, however, if the suggested rating forms actually do prove useful to maintenance management and do prove to be statistically valid as a measurement of performance. In any case, the suggested rating forms should provide useful maintenance personnel performance data for use in developing a model which accurately explains the contri-

bution of individual motivation and ability to Air Force
maintenance.

Chapter 3

METHODOLOGY

This chapter deals with the procedure used to test the recommended performance measures of quantity and quality. In addition to the test utilizing maintenance personnel at Williams AFB, Arizona, a small, independent survey was distributed to determine the opinions of several maintenance managers at other bases on the usefulness of the performance measures.

The following topics will be considered in Chapter 3:

1. The data gathering instruments.
2. The sample selection procedures.
3. The specific data collection procedures.
4. The plans for analysis of the data.

A discussion of assumptions and limitations of the methodology will be included in the chapter summary.

The Data Gathering Instruments

The appraisal forms developed in Chapter 2 (Appendices B and C) are the primary data sources for this study. Based on the conclusions of the previous chapter, these forms were developed with the intention of providing useful performance information for statistical analysis.

In addition to the appraisal forms, a limited number of maintenance officer opinions were solicited concerning

the validity of the proposed appraisal methods. It was felt that a general survey of maintenance officers would have been overly time-consuming and costly. The overall results of such a survey would have revealed a consensus opinion of average officers, whereas the opinions of five or six officers who have excellent performance records may be considered more relevant. An example of the officer opinion questionnaire and its cover letter are contained in Appendix K. For such a small survey, free form answers were solicited rather than multiple-choice or two-way answers; this format was suggested by Neter and Wasserman (1973:27). The first question on the survey and the cover letter were designed to establish rapport with the respondent. All questions were simple and were designed to be clear, to avoid leading the respondent, and to eliminate bias. The questions included the following:

1. Is individual performance important to the maintenance organization?
2. Are the ranking forms appropriate for appraising performance or can you suggest a better approach?
3. Are quantity and quality useful measures of performance?
4. Which do you consider to be more important, quantity or quality?
5. If one is more important than the other, can you indicate how much more important it is?

The responses to this survey questionnaire were intended to

indicate whether the proposed performance rating forms have face validity and how quantity and quality ratings might be combined into a single measure of performance.

Several existing sources could theoretically provide data to validate the accuracy of the performance appraisals. Among these are Airmen Performance Ratings (APRs) and Skill Knowledge Tests (SKTs). However, both the APRs and SKT results are Air Force privileged information and were not available to this author as data sources. Further, APRs have a history of being highly inflated, according to Callander (1970:4), while SKTs are paper and pencil tests administered only to test selected skills and at uneven intervals. Thus, neither would in actuality be an appropriate source for validating the performance ratings. In fact, there are no available Air Force records that would be useful for validation of speed of performance--or quantity--rating data.

The source used in this research for a possible validation of at least the quality of performance rating scale may be referred to as maintenance quality control (QC) information on technician inspections under the Air Force Maintenance Standardization and Evaluation Program (MSEP). MSEP personnel performance scores are based on separate failure levels or baselines for each type of maintenance task. Although they provide the best existing performance data for validation use, difficulties may be encountered in using MSEP. For instance, personnel evaluations are not

required nor completed for any one individual on any regular basis; thus some of the technicians in any random sample of technicians may not have MSEP records.

Summary of Data Needs

The data to be gathered for this research thus includes (1) supervisor appraisals of technicians reporting to them, using the recommended rating forms (Appendices B and C), (2) surveys of a few selected maintenance officers (Appendix K), and (3) MSEP reports available for most of the technicians drawn in the sample. The sample instruments for (1) above were reviewed and approved by the Air Force Military Personnel Center, Randolph AFB, Texas. Human subject clearances and a privacy statement example are included in Appendix A.

Sample Selection Procedures

Selection of the maintenance technicians for the sample proved to be difficult. It was desired that only line technicians be evaluated, and that they be evaluated by their immediate supervisors, who are responsible for scheduling and inspecting assigned tasks. No existing source document used by the Air Force appears to reflect this information on an accurate and current basis. A complete listing of all maintenance line supervisors and their immediate subordinates is necessary if a randomly drawn sample of shift supervisors and their subordinates is to accurately reflect the entire

maintenance organization of a chosen Air Force Base.

The existing personnel listings of rating officials responsible for preparation of annual APRs does not reflect current work group assignments. For instance, in examining the personnel listing at Williams AFB, Arizona, it was found that one supervisor was shown with six subordinates, none of whom currently reported to him. Another supervisor on this same personnel listing had three assigned subordinates listed, one of whom was not currently assigned to him, while he actually supervised an additional seven who did not appear on the listing. As a consequence, the existing listing of rating officials and subordinates could not be used in selecting the sample.

Another data base, the Maintenance Management Information and Control System (MMICS), maintains a file of all personnel assigned to the maintenance organization, but the file does not identify supervisors or their immediate subordinates. As a consequence, it was necessary to obtain a current roster from each maintenance section prior to drawing the technician sample for the study.

In this study supervisory groups were randomly drawn from a listing of all such groups in each maintenance squadron, by using the random number tables in Beyer (1978:544). In a few cases supervisory groups were also chosen by dice rolls to decide the particular shifts to be included in the sample. At least three and no more than five subordinates were selected to be evaluated by each supervisor.

Where more than five subordinates reported to one supervisor, five subordinates were selected for the study by again using random number tables. Alternates were also selected by this method where more than five technicians were encountered in a group. This stratified selection method allowed the researchers to control sample participation, to eliminate supervisor bias, to obtain a representative sample, and to allow supervisors to evaluate enough subordinates (five each) to obtain valid comparisons.

Sample size was set at ninety technicians per squadron (approximately twenty percent of the population) based on the sample sizes used in similar studies. Eighteen corresponding supervisors per squadron completed technician evaluations. As the maintenance organization at Williams AFB is made up of two squadrons, the total sample drawn consisted of 169 technicians (some supervisory groups had less than five technicians) and thirty-six supervisors. No research dealing with Air Force maintenance technician performance has reported useful information on the effect of sample size on statistical tests. This research should provide information on adequate sample sizes for minimizing the probability of erroneously accepting a hypothesis (Type II error) for certain statistical tests.

As for the maintenance officers selected to answer the independent survey questionnaire, they are personal acquaintances of the author and have all managed maintenance

operations graded as excellent or better by Major Command inspectors. The majority are now retired and should thus have felt no restrictions in supplying candid answers.

As has been noted, sample selection for this study was very time-consuming. To begin with, much time was spent in evaluating existing rosters of personnel before it was determined that the rosters were inadequate for drawing the names of maintenance personnel and their immediate shift supervisors. The sample was randomly drawn from all supervisory groups with more than three technicians reporting to a shift supervisor. Details of the exact data collection process follow.

Data Collection

The sample data was collected at Williams AFB, Arizona. To protect the privileged nature of the supervisor performance appraisals and Privacy Act requirements, control numbers were assigned to the technician participants in the study and the researchers supervised all appraisal and evaluation sessions. All participants completed the survey forms in a central location and during specified time periods which allowed for participation by personnel from all three shifts.

The independent officer survey questionnaires were mailed to eight maintenance officers and responses were received from four, as the remainder had moved and left no forwarding addresses.

In Chapter 2 it was theorized that peer appraisals might deflate ratings; it was also noted that no research had been done in this area. No peer appraisals were attempted in this study either, due to time constraints, to the difficulty of providing the necessary information to all 169 participants, and to the wide range of experience levels found among any five technicians within a supervisory group. The sample size simply made peer appraisals too difficult to administer, which may also be the reason why this method had not been used in the past.

Data collection was spread over a period of two-and-one-half weeks. Unfortunately, during the survey period several technicians changed shifts and supervisors. The collection time period also allowed for changes in work requirements. These difficulties were relatively minor, however, and the data collection methods were successful.

Data Analysis

In analyzing the data obtained from the Williams AFB sample, the first consideration is to determine if the organization is adequately represented in the sample. In this case, the maintenance organization is small enough to compare the sample population and the base maintenance population with respect to several characteristics.

Statistical analysis of the data should then establish if it is suitable for use in a regression analysis and if there are significant differences between quantity and

54

quality ratings. Next, comparisons will also determine if quantity is related to quality based on the supervisors' subjective appraisals of the technicians. Finally, an attempt to validate the quality rating scale will use correlation analysis to find if any linear relation exists between MSEF data and the performance quality data of this study.

Descriptive statistics will also be generated using the Biomedical (BMD) Detailed Data Description program (1977) and the Bivariate Plot (POD) program as described by Dixon (1977). The decision to use BMDP was based on the researcher's familiarity with the program and on a preanalysis by Dunn and Francis (1978:65) and Muller (1978:71).

Finally, the maintenance officer survey responses will be summarized and a consensus opinion, if there is one, will be reached concerning the relative importance of quality versus quantity of performance in maintenance.

Summary

This chapter has covered the methodology involved in obtaining the data for this study: the data gathering instruments, sample selection, data collection, and the data analysis plans. At this point of the analysis, obvious restrictions and limitations are few, but some do exist. For example, uncertainty exists regarding the error produced by the specified sample size, a difficulty which can be removed only after analysis is complete. In addition, the

small number of maintenance officers surveyed concerning the relevance and usefulness of the appraisal forms may produce results of limited use. Finally, the lengthy data gathering period of two-and-one-half weeks allowed for a number of changes in personnel and policy which may affect results. Most of these limitations will be resolved in the next chapter on analysis.

Chapter 4

ANALYSIS

The analytic purpose of this study is to develop a subjective performance rating method that provides the following:

1. Performance data applicable to regression analysis as a dependent (Y) variable.
2. Performance data that has some apparent or actual validity compared to existing performance measures.
3. Performance data with desirable statistical properties.
4. Performance data that accurately reflects the organizational composition.

The analyses of the sample data that will be discussed in this chapter are based on the above requirements. First, the sample will be analyzed to determine if it reflects organizational composition. Secondly, the sample data's statistical properties will be considered. Thirdly, the quantity and quality ratings will be compared. Fourthly, the association between existing performance measures (MEEP) and the sample quality ratings will be investigated. Finally, the maintenance officer responses to the opinion survey will be summarized.

Sample Composition

All of the technicians and supervisors included in the sample were members of the maintenance organization at Williams AFB, Arizona, a USAF pilot training base. This maintenance organization differs from most base organizations in that the avionics repair function is a branch of the Field Maintenance Squadron (FMS) and not a separate squadron. The Williams AFB organization structure, depicted in Figures 8 through 13, is otherwise comparable to the general Air Force organization structure depicted in Figures 1 through 5. The actual number of line technicians was extracted from the Maintenance Management Information and Control System (MMICS); these figures may differ from authorized strength limits and, since supervisors are excluded, may not coincide with squadron strength figures.

The random sample of 169 technicians was intended to represent the entire maintenance organization. Eighty technicians were selected from the Operational Maintenance Squadron (OMS), or 22.6 percent of the squadron. Eighty-nine technicians were selected from the Field Maintenance Squadron (FMS), or 18.7 percent of the squadron. The grade distributions of the sample closely parallel that found in the organization (see Table 4). The FMS sample and squadron are composed primarily of sergeants and civilians, while OMS is primarily composed of airmen.

The relative representation of squadron branches in

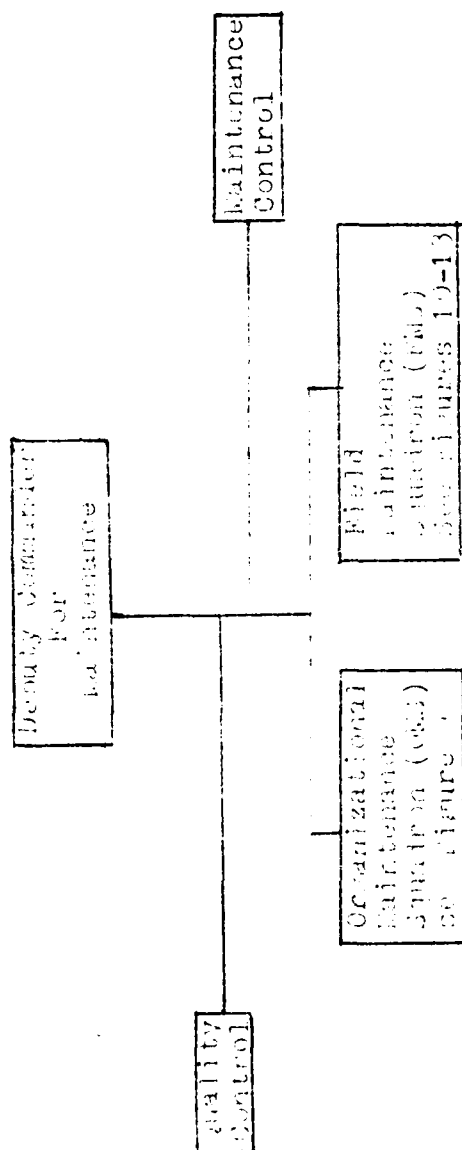


Figure 3

Deputy Commander for Maintenance Organization
 Williams AFB, Arizona
 Sources: Unit Manning Document and AFB.

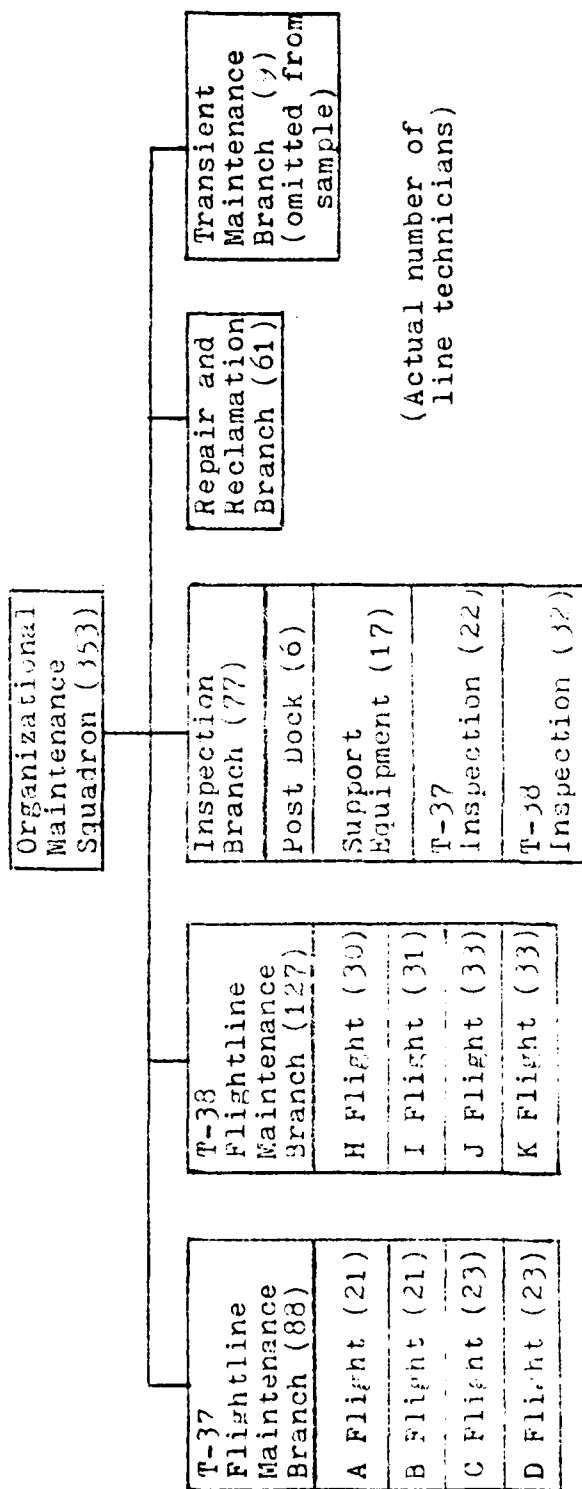
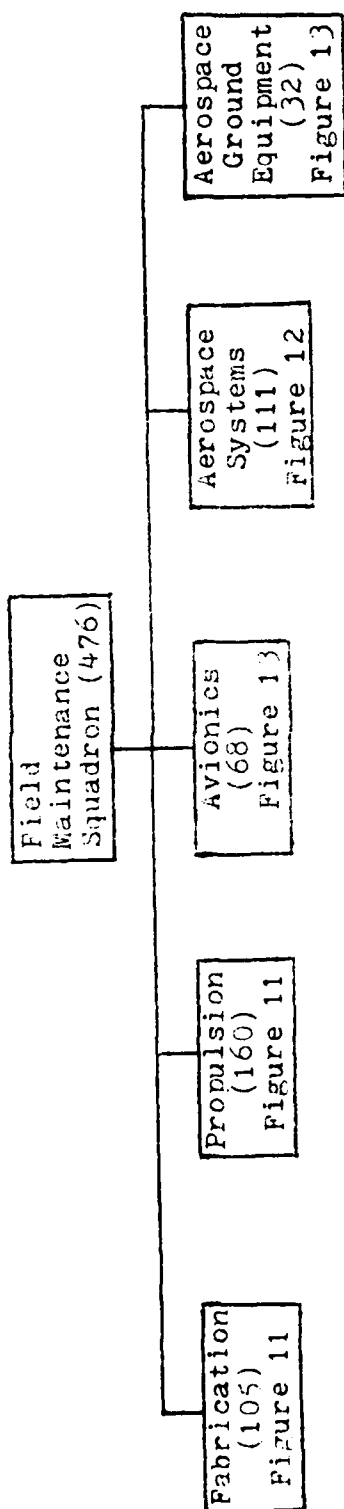


Figure 9

Organizational Maintenance Squadron Organization
Sources: MELCS and Unit Manning Document



(Actual number of
line technicians)

Figure 10

Field Maintenance Squadron Organization
Sources: LMICU and Unit Manning Document

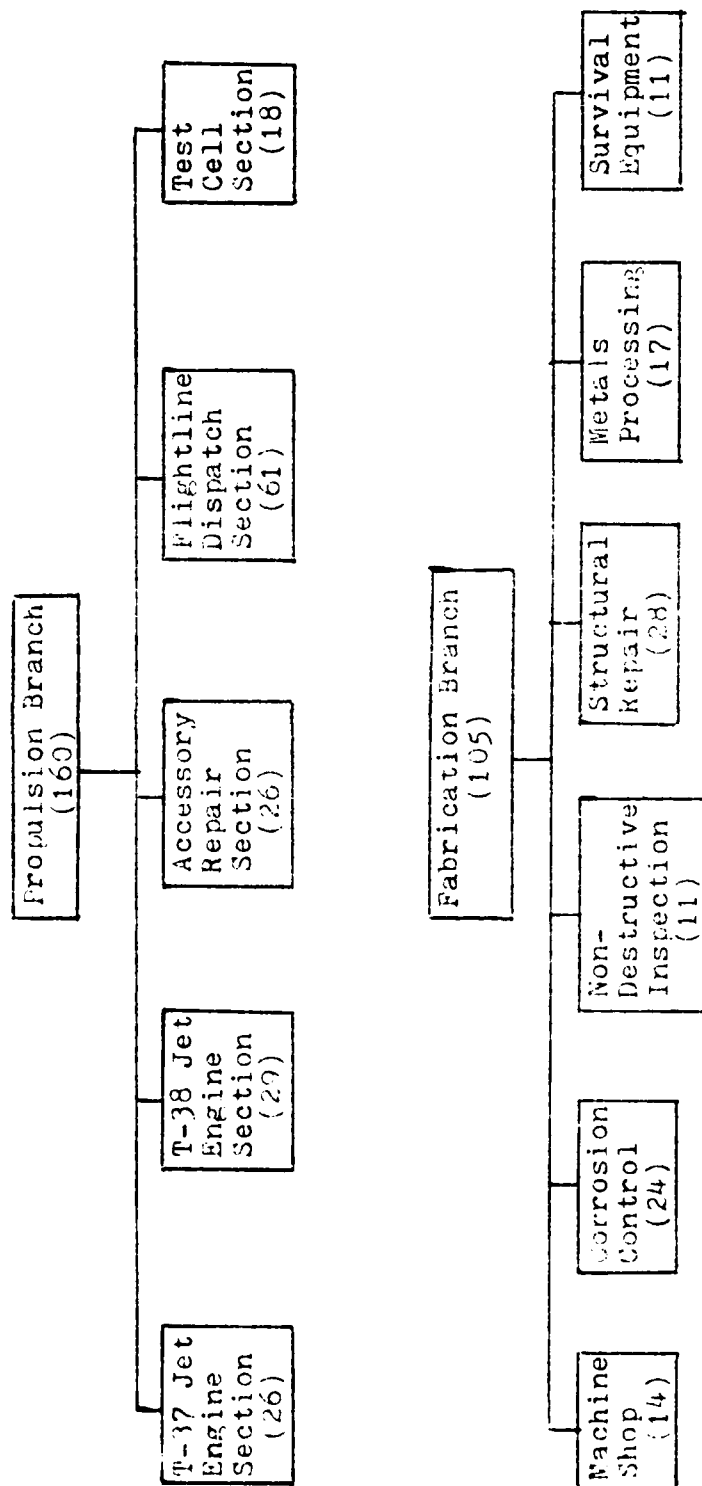


Figure 11

Propulsion and Fabrication Branches, FAS
Sources: EMICS and Unit Manning Document
(Actual number of line technicians)

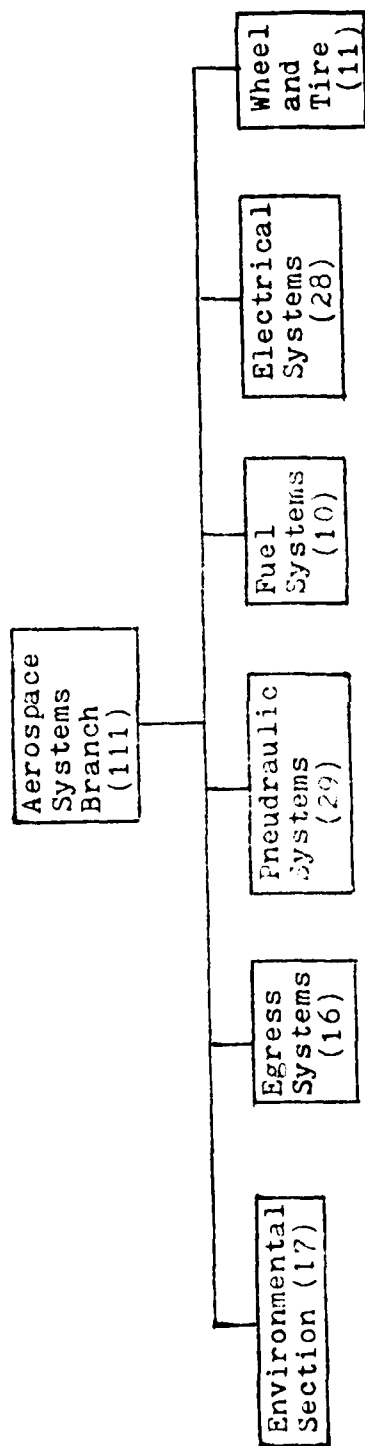


Figure 12

Aerospace Systems Branch, FMS
Sources: FMCS and Unit Manning Document
(Actual number of line technicians)

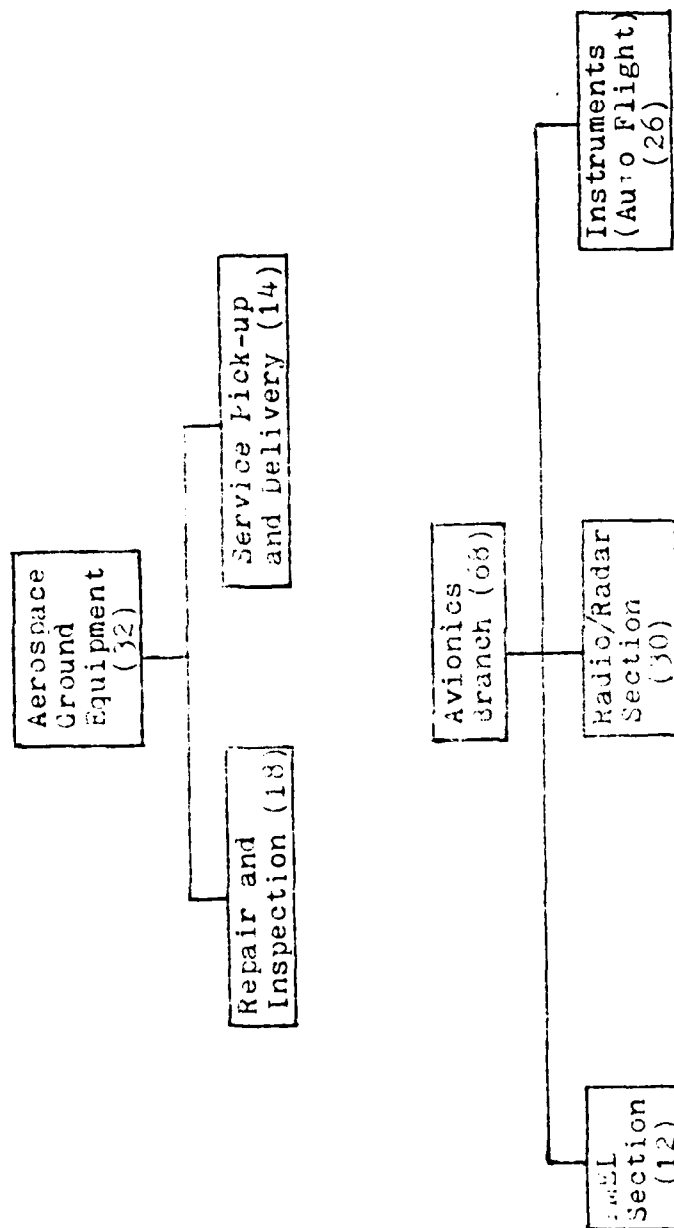


Figure 13

AGE and Avionics Branches, FMS
Sources: FMS and Unit Manning Document
(Actual number of line technicians)

the sample is depicted in Table 5. The OMS sample closely parallels the squadron population. The FMS sample is not as representative due to the large number of work sections with few technicians.

The OMS sample included eight day shift supervisory groups, eight swing shift groups, and two mid (i.e., graveyard) shift groups. The FMS sample included twelve day shift groups and six swing shift groups. FMS had few swing shift groups and few mid shift groups with three or more technicians assigned.

Table 4

Grade Distribution of Sample and Population

Organ.	Total	Civ.	TSgt	SSgt	Sgt	A1C	ASN	AS
<u>FMS</u>								
Population	476	122	20	53	76	175	13	17
Percentage	100%	26%	4%	11%	16%	37%	3%	4%
Sample	89	23	3	11	16	29	0	1
Percentage	100%	32%	3%	13%	18%	33%	0%	1%
<u>OMS</u>								
Population	336	18	3	23	46	206	34	25
Percentage	100%	5%	1%	7%	13%	58%	9%	7%
Sample	30	4	0	4	7	54	3	5
Percentage	100%	5%	0%	5%	9%	68%	3%	5%

(population information from MMICS)

In short, the sample appears to represent fairly the maintenance organization as a whole. The random selection of supervisory groups with a subsequent random selection of no more than five technicians from each group appears to have

provided a representative sample in terms of rank structure and experience level and relative branch strengths.

Table 5

Relative Squadron Branch Strengths
versus Relative Sample Branch Strengths

Organization	Population	% of Squadron	Sample	% of Sample
<u>CMS Branches</u>				
T-37 Flight	33	25.6	23	26.1
T-38 Flight	127	36.6	23	26.6
T-37 Inspect.	22	6.1	6	6.9
T-38 Inspect.	32	9.6	8	11.1
Port Inspect.	6	2.3	5	5.56
Rep. and Reclamation	61	17.6	10	13.6
Support Inspect.	17	5.6	5	6.61
Total	<u>253</u>		<u>50</u>	
<u>FMS Branches</u>				
Propulsion	160	33.6	20	22.4
Fabrication	105	22.5%	17	32.8
Aerospace Systems	111	23.3%	30	33.6
AME	32	6.7%	5	5.6
Avionics	63	14.2%	5	5.6
Total	<u>470</u>		<u>57</u>	

(population information from AMICS)

Sample Statistical Properties

All technicians included in the random sample were rated on quality and quantity of performance by their immediate supervisors, using the recommended forms in Appendices B and C. Only the data from the rating scale was used; the rankings were intended to aid the supervisors in their rating efforts. The data collected is shown in Appendix D.

The rating forms were originally designed to allow for continuous ratings between zero and ten. This would have produced histograms with small cell widths using, for instance, a scoring procedure where a 9.0 score or rating value resulted from a rating between 8.75 and 9.25. In this sample most ratings were grouped immediately around the rating numbers, which required adjustment to the following rating system:

<u>Rating Range</u>	<u>Rating Value</u>
> 9.5	10.0
> 8.5-9.5	9.0
> 7.5-8.5	8.0
⋮	⋮
> 0.5-1.5	1.0
0.5	0.0

Use of this rating system resulted in the quantity and quality rating values reported in Appendix D.

The resulting histograms for quantity and quality of

performance for OMS and FMS are presented in Figures 14 and 15. According to Bradley (1968:55), the discrete distribution case can be treated in the same way as the continuous distribution case. Difficulties arise in the use of non-parametric tests due to the large number of equal values and the potential loss of test power.

The sample means and standard deviations were calculated using the BMD Detailed Data Description program (P2D). This information is included in Appendix E. This program also computed values for skewness and kurtosis, which will be discussed following an analysis of the MSEP data.

MSEP Data

The Maintenance Standardization and Evaluation Program (MSEP) data is also included in Appendix D. This is the existing information with which quality ratings are to be compared in an attempt to establish some validity for the subjective quality ratings. The MSEP data represents two years of personnel inspections while the subjective ratings represent supervisors' appraisals at one particular point in time. The MSEP personnel evaluations are based on compliance to equipment specifications; failure of an evaluation represents either a major safety discrepancy or the accumulation of more than the allowed number of minor discrepancies for a particular task. Inspections take into consideration completed maintenance actions (CMA), completed maintenance inspections (CMI), and task evaluations (TE).

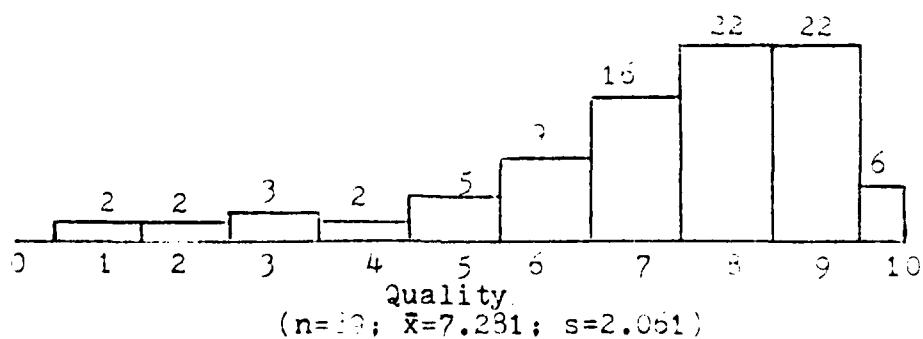
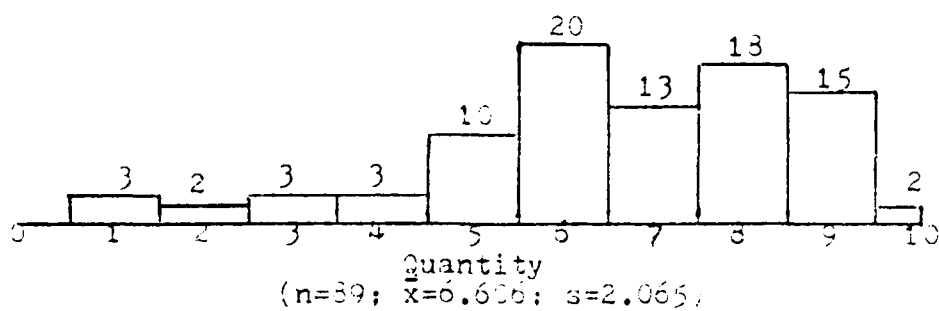


Figure 14

FMS Histograms

(Statistics from Appendix E. BMDP analysis; the number of technicians recorded in each interval is noted at the top of each frequency column)

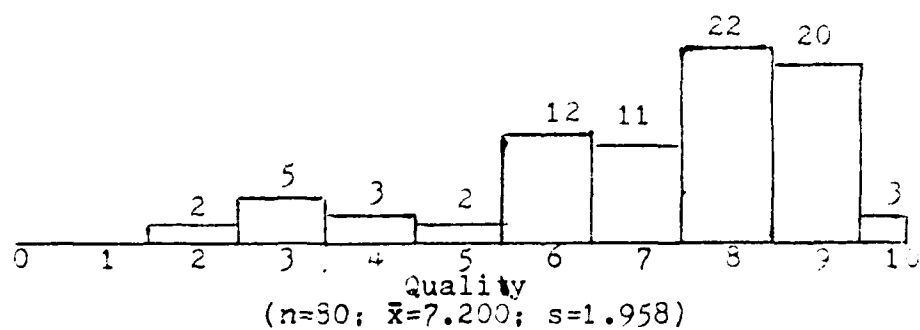
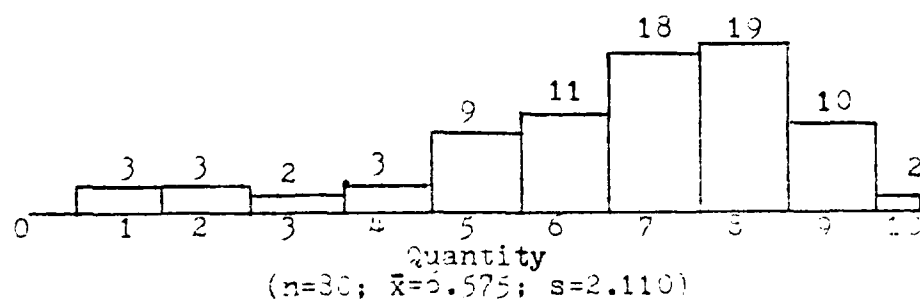


Figure 15

OMS Histograms

(Statistics from Appendix, BMDP analysis; the number of technicians recorded in each interval is noted at the top of each frequency column)

Since some supervisors also work as line technicians, completed supervisory inspections (CSI) and supervisor evaluation (SE) are also included. The MSEP score used in this analysis is based on a weighted average of all inspections subtracted from one so that an individual MSEP score of 1.0 indicates that no discrepancies were noted during an inspection of a particular technician's work.

The management of the maintenance organization at Williams AFB requires that every technician receive an inspection every eighteen months, although the Air Force does not require an inspection so often. In spite of this policy, a large proportion of the technicians from both squadrons were relatively new arrivals and had not been inspected (see Table 6). On this basis alone, the MSEP data

Table 6

Ratio of Personnel Awaiting
MSEP Action/Total Personnel (Maintenance)

Organization	Number	Percentage
OMS Sample	17/80	21.25
OMS Population*	140/542	25.83
FMS Sample	19/89	21.35
FMS Population*	136/454	29.96
* Population data is from MDCS and includes all supervisors and technicians.		

would be incomplete and unuseable as a performance measure for the entire organization. It is also significant to note

that of those technicians inspected who were included in the sample, OMS technicians received an average of 2.7 inspections in the past two years, compared to 1.7 inspections in the past two years for FMS technicians.

The histograms of the quality ratings and MSEP scores for those technicians in the sample who received inspections are presented in Figures 16 and 17. The data concerning sample means and standard deviations was also obtained using the BMDP program. A preliminary inspection of the histograms might indicate that the higher average number of inspections of technicians in OMS deflated OMS MSEP ratings in comparison to FMS MSEP ratings.

Based on the histograms presented in this section, the quality and quantity of performance rating data will now be analyzed with regard to important statistical qualities. Most significant is the relation between the histogram distributions and the normal distribution.

Normality and Applicability to Regression Analysis

The maintenance technician ratings were designed for use in regression analysis as the dependent (Y) variables. One of the assumptions of multiple linear regression is that the observations for the dependent variables are independent and drawn from a normal distribution (Neter and Wasserman; 1974:219). Normality is also required for most parametric tests that could be used to compare the quantity

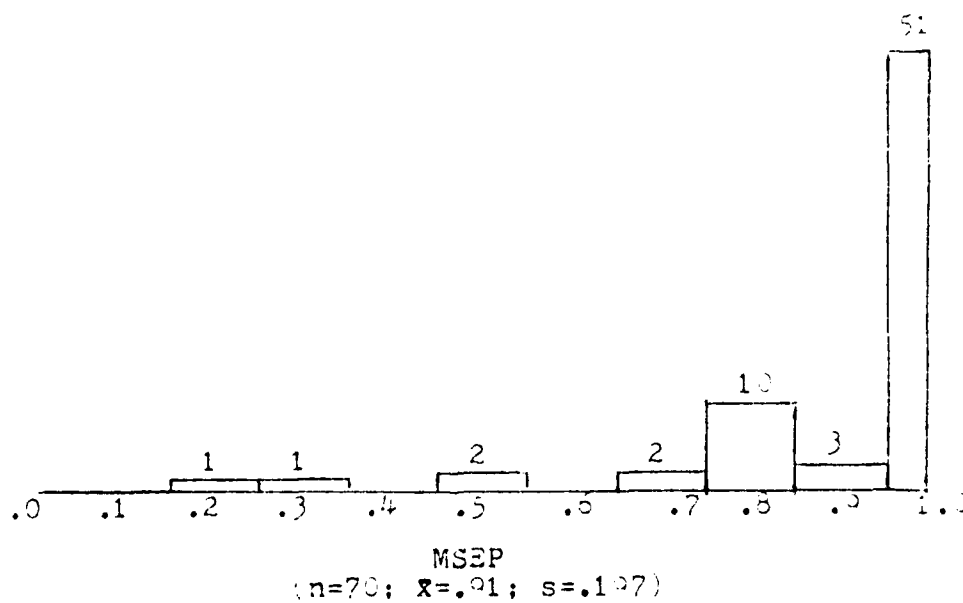
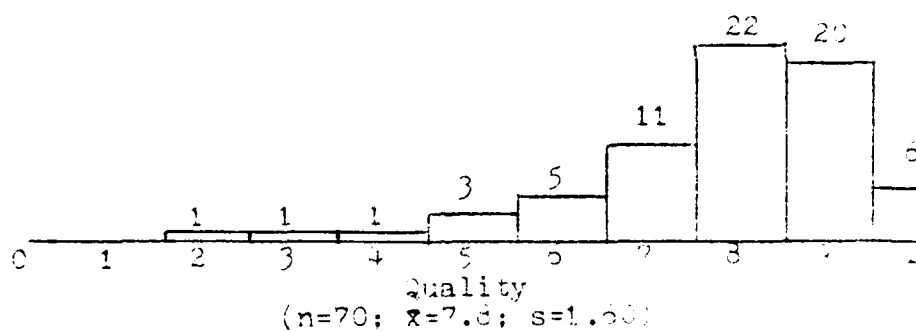


Figure 16

FMS Quality Ratings and MSEP

(Statistics from Appendix J, Correlation Analysis: the number of technicians recorded in each interval is noted at the top of each frequency column)

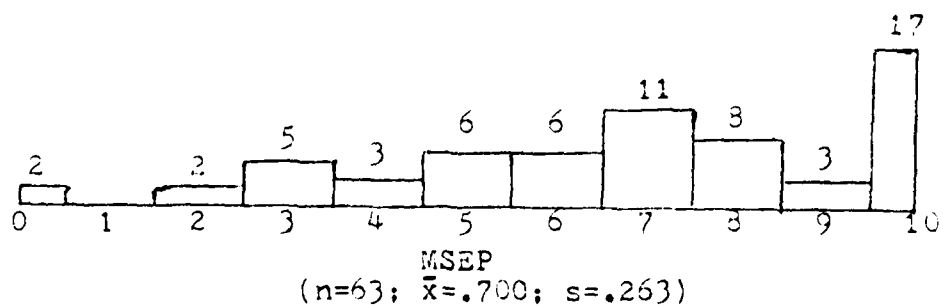
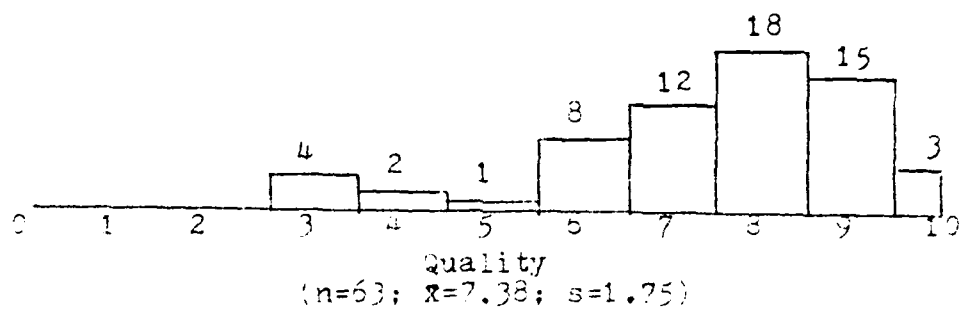


Figure 17

OMS Quality Ratings and MSEP

(Statistics from Appendix J, Correlation Analysis; the number of technicians recorded in each interval is noted at the top of each frequency column)

and quality histograms.

Tests which can be used to test the normality assumption include the Kolmogorov-Smirnov and Chi-Square goodness-of-fit tests. The Kolmogorov-Smirnov test is not applicable in this case, as population parameter estimators cannot be specified in advance of testing (Lapin;1973:426). The Chi-Square test does allow the use of sample estimators for testing population parameters. The Chi-Square test, however, is more difficult to interpret and the probability of accepting a false null hypothesis (Type II error) is not well defined (there are several ways in which the null hypothesis could be false). Calculations for the Chi-Square goodness-of-fit test for the RMS quality of performance data are included in Appendix F. A summary of the Chi-Square results for all of the sample data is included in Table 7.

The small Type I error significance levels (α) indicate poor fits of all the data to the normal distribution. For instance, the .01 level of significance for the RMS quality data indicates that if a distribution were normal with a mean of 7.2 and a standard deviation of 2.0, then there would be a 1 percent chance that a sample could be obtained from this distribution yielding a Chi-Square value equal to or greater than 5.96. Similar interpretations of the remaining Type I error levels indicate that all the sample data have relatively low probabilities of being from normal distributions.

The relatively poor fit of the sample data to normal

Table 7
Chi-Square Normal Distribution
Goodness-of-Fit Tests

Sample	Distribution Fit to $N(\bar{x}; s^2)$	χ^2	ν^a	Type I ^b Risk	n
OMS; Quality	$N(7.2; 2.0^2)$	5.96	1	.01	30
OMS; Quantity	$N(6.6; 2.1^2)$	4.60	1	.02	30
OMS; MSEP	$N(.70; .26^2)$	9.79	2	.001	63
OMS; Rev. MSEP ^c	$N(.66; .24^2)$	2.73	2	.25	52
FMS; Quality	$N(7.3; 2.1^2)$	17.98 ^d	4	.001	30
FMS; Quantity	$N(6.6; 2.1^2)$	6.41	2	.02	30
FMS; MSEP	$N(.91; .20^2)$	Too inflated to calculate			

^a ν =degrees of freedom.

^bType I risk is the alpha level of significance.

^cRev. MSEP includes inspection results for technicians with two or more inspections.

^dSee Appendix F for Chi-Square calculations.

distributions, fortunately, has minimal effect on the multiple linear regression model with which the data will be used. The regression coefficient parameter estimators will remain unbiased and consistent though not highly efficient. If the lack of normality is significant enough, the error terms (residuals) of the regression model may not have constant variance (a condition defined as heteroscedasticity). This condition can best be investigated after a preliminary regression model is developed and, if non-constant variance results, a transformation of the performance data can be made to correct the problem. For instance,

if the error term standard deviation is proportional to the square of the factor level mean, the reciprocal transformation stabilizes variances (Neter and Wasserman;1974:507).

It is often the case that the same transformation which helps stabilize the variance also helps normalize error terms.

Neter and Wasserman (1974:123) state that:

It is therefore desirable that the transformation for stabilizing the error variances be utilized first, and then the residuals studied to see if serious departures from normality are still present.

It is thus apparent that normality of the sample distributions is not a serious initial concern for regression analysis. Non-normality does have serious effects on tests which can differentiate between the mean and variance of the sample distributions. The skewed nature of the distributions appears to be the particular quality leading to the poor comparison of the sample distributions to normal distributions. All of the sample distributions are skewed as indicated by the data in Table 3. The effect of skewed departure

Table 3.

Skewness and Kurtosis of Sample Data

	OMS		FMS	
	Quantity	Quality	Quantity	Quality
Skewness (δ_1^1)	-.254	-1.009	-.314	-1.212
(δ_1^2)	.210	1.213	.553	1.469
Kurtosis (δ_2^2)	.462	.251	.372	1.144
(data from BMDP analysis, Appendix E.)				

from normality on parametric tests to compare means is

significant for small sample sizes but insignificant for the large sample sizes used in this study (Davies;1956:55). Most parametric tests for the comparison of means, however, require equal sample variances. And tests that determine variance equality are significantly affected by skewed sample distributions. The sample distributions all, with one exception, have equal variances with a Type I risk of five percent (Table 9). A calculation to test for the equality of variances for the FMS quantity and quality data is included in Appendix H. However, according to Davies (1956:55), with the degree of skewness and kurtosis exhibited by this data the Type I risk becomes closer to 6.8 percent (see Appendix G). This discrepancy does not decrease as sample size is increased.

In summary, it is apparent that the sample distributions exhibit marginal normality due to skewness. As the normality assumption is not critical in the use of the multiple linear regression model, this data is applicable to regression but should be used with caution. In particular, the regression residuals should be carefully inspected for heteroscedasticity and, if necessary, remedial transformations should be utilized. The use of parametric tests for the comparison of sample distributions is questionable due to the skewed nature of the distributions. In particular, it is difficult to test for the equality of variances with a Type I error of less than seven percent, while variance equality is a requirement for most parametric tests. Non-parametric

Table 9
Comparison of Variance Equality
(see Appendix H)

Variance 1	Variance 2	F	ν_1, ν_2^a	Type I Risk
FMS Quant. (4.25)	FMS Qual. (4.26)	.9966	88,88	.05
FMS Quant. (3.91)	FMS MSEP (2.57)	1.520	69,69	.05
FMS Quant. (6.20)	FMS RMSEP ^b (3.32)	1.867	35,35	.05
OMS Quant. (4.45)	OMS Qual. (3.83)	1.174	79,79	.05
OMS Qual. (3.08)	OMS MSEP (6.93)	.444	62,62	N.S. ^c
OMS Qual. (3.35)	OMS RMSEP ^b (5.52)	.508	51,51	.05
OMS Quant. (4.26)	FMS Quant. (4.45)	.958	88,79	.05
OMS Qual. (4.25)	FMS Qual. (3.83)	1.109	88,79	.05

^a ν =degrees of freedom.

^bRMSEP includes inspection results for technicians with two or more inspections.

^cN.S. (Not Significant). The null hypothesis is rejected and the variances cannot be accepted as equal.

tests may be applicable if the sample histogram distributions can be used with this type of test.

Rating Comparisons

The next area of interest is the comparison of the quantity and quality ratings within and between squadrons to determine if any differences exist for the rating distributions. Due to the non-normal distributions, the tests for differences will be conservative and will either be non-parametric or parametric, with variance equality not required.

According to Brailley (1968), Gibbons (1976), and Hollander and Wolfe (1973), most non-parametric tests are based on ranking procedures. Validity is seriously affected by a large number of ties in the sample data, and that occurs for this study data. One non-parametric test that is applicable for the comparison of sample distributions is the Kolmogorov-Smirnov two sample test (Gibbons:1976:252). For this test only two assumptions are needed relative to the study data: first, that the quantity and quality ratings for each technician by a single supervisor are independent, and, secondly, that the quantity and quality of performance data be considered as continuous variables. Thus the probability (P) values for this test should be considered to be conservative (Gibbon:1976:258). The Kolmogorov-Smirnov test results summarized in Table 10 indicate that the quantity and quality rating distributions for each squadron

are different; however, the test does not specify in what way the distributions are different.

To compare means for the different sample distributions, the Behrens-Fisher approximate "t" test was used. This test does not require equal variances or equal sample size, although the formulation and solution are somewhat controversial according to Dudewicz (1976:311). The results of the approximate "t" tests are summarized in Table 10. These results indicate a significant difference between the means of the quantity and quality distributions within the squadrons. No difference, however, is noted between the means of the distributions between the squadrons. A comparison of the means for squadron quality of performance distributions and MSFP appraisals indicates a difference for FMS and no difference for CMS. In the cases where significant differences were found, the probability of erroneously accepting the alternate hypothesis (Type II error) for the difference in means tested ranged from zero percent to thirty percent. The higher Type II errors were found on the tests comparing the means of the quality and quantity performance distributions within squadrons. It appears that the Type II error size is influenced by the histogram cell width producing a large sample variance.

In summary, conservative tests indicate significant differences in the quantity and quality distributions within the CMS and the FMS squadrons, but not between squadrons. This significant difference in the distributions does not

Table 10
Comparison of Means
(see Appendix I)

Mean	Mean	Test ^a	Stat. ^b	ν^c	Sig. ^e Type I Risk	Type II Risk
FMS Quant.(6.6)	FMS Qual.(7.3)	K-S B-F	.20(D ₊) 2.2(t)	- 175	.05(P) .05(α)	- .30(β)
OMS Quant.(6.57)	OMS Qual.(7.2)	K-S B-F	.17(D ₊) 1.9(t)	- 155	.10(P) .10(α)	- .13(β)
OMS Quant.(6.57)	FMS Quant.(6.6)	B-F	.57(t)	164	None(α)	-
OMS Qual.(7.2)	FMS Qual.(7.3)	B-F	1.4(t)	166	None(α)	-
OMS Qual.(7.2)	OMS ^d MSEP (7.9)	B-F	.50(t)	111	None(α)	-
FMS Qual.(7.3)	FMS ^d MSEP (7.1)	B-F	5.6(t)	151	.05(α)	.01(β)

^aTests; K-S, Kolmogorov-Smirnov Non-Parametric two sample test; B-F, Behrens-Fisher Approximate "t" test

^bCalculated test statistics; D₊ for the Kolmogorov-Smirnov test, t for the Behrens-Fisher Approximate "t" test.

^c ν =degrees of freedom.

^dThe MSEP means were multiplied by ten for these tests.

^eSignificant Type I risk levels were reported when a significant difference in means was found.

necessarily mean that individuals will have significantly different quantity and quality ratings. The significant difference does indicate that both distributions should be used in the regression analysis. The association between ratings and the validity of the quality of performance ratings will be considered next.

Rating Associations

To determine the degree of association between quality and quantity ratings and between quantity ratings and MBSF data requires some sort of association measure. Non-parametric measures would be preferred in this case since the parametric measure or linear correlation coefficient assumes that the distributions are drawn from a bivariate normal distribution (Neter and Wasserman;1974:394). Since all the sample distributions are skewed and provide poor fits to the normal distribution, it is unlikely that the basic assumption for the use of the correlation coefficient can be met. However, non-parametric association measures require that the data be ranked, whereas the sample data in this study contains too many ties for such measures to be valid.

Since non-parametric measures cannot be used, it is necessary to use the Pearson product-moment correlation coefficient as a descriptive measure. In this regard, Freund (1977:324) states the following:

Note that the sample correlation coefficient r is often used to measure the strength of a linear relationship exhibited by sample data even if the

data do not come from a bivariate normal population.

Gibbons (1976:339) indicates that the use of an interval scale, as is used in the ratings under study, to subjectively measure performance also allows for the use of the Pearson product-moment correlation coefficient as a descriptive measure.

In this study all correlation coefficients were calculated using the BMDP program. The results of these comparisons are summarized in Table 11. The significance levels reported in Table 11 are based on a test to determine if the coefficients are equal to zero or not; significant results indicate non-zero coefficients. These results are limited by the fact that it cannot be shown that the sample distributions are drawn from a bivariate normal distribution for the population, a basic assumption for the use of the correlation coefficients.

Table 11

Summary of Correlation Coefficients (r)
for Quantity/Quality Associations and
Quality/MSEP Associations
(see Appendix J)

	Quant.		MSEP		Rev. MSEP ^c	
	n	r	n	r	n	r
OVS Qual.	80	.821 ^a	63	.186 ^b	51	.603 ^a
FMS Qual.	89	.835 ^a	70	.181 ^b	50	.687 ^a

^aCorrelation significantly non-zero at the .01 level.

^bCorrelation is not significantly different from zero.

^cRev. MSEP includes personnel with two or more incorrect.

Interpretation of the correlation coefficients as descriptive statistics indicates that quantity and quality of performance ratings for individual technicians are highly correlated. There does not appear to be any linear correlation, however, between technician quality ratings and MSEP inspections. If, however, only those individuals who have had two or more inspections in the previous two years are included (revised MSEP), a significant correlation results for OMS personnel. However, there is still no correlation between FMS ratings and MSEP data.

It thus appears that the quantity and quality ratings for individual technicians are highly correlated. Very poor correlation exists between quality ratings and MSEP data. These correlation statistics should be considered as descriptive statistics only.

This section concludes the analysis of the sample data from the performance ratings. Next, the utility of these ratings will be considered based on the observations of a few maintenance officers.

Opinion Survey Analysis

The opinions of a limited number of maintenance officers were solicited concerning the usefulness of the performance ratings and the relative importance of quantity and quality of performance. The survey questions and letter of transmittal are contained in Appendix K. Survey questionnaires were mailed to eight maintenance officers the prior

AD-A090 635

AIR FORCE INST OF TECH WRIGHT-PATTERSON AFB OH
AIR FORCE MAINTENANCE TECHNICIAN PERFORMANCE MEASUREMENT, (U)
DEC 79 J R HICKMAN
AFIT-CI-79-241T

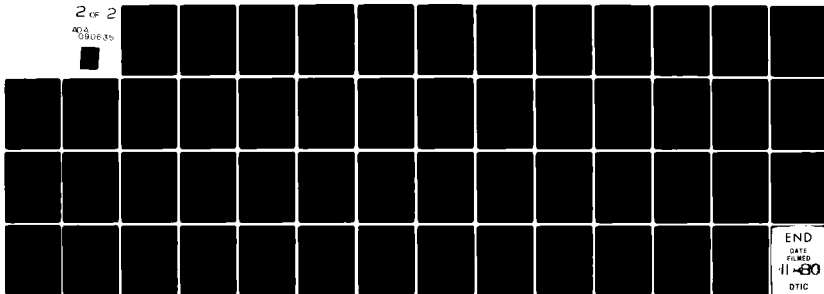
F/G 5/9

UNCLASSIFIED

NL

2 of 2

AD-A
090635



END
DATE
FILMED
41-40
DTIC

knows personally; replies were received from four, as the remaining four had moved and left no forwarding addresses. The Deputy Commander for maintenance at Williams AFB also responded to the questionnaire. In all, one Major, one Lieutenant Colonel, and three Colonels responded. The responses are compiled in Appendix L and represent three maintenance organization commanders, one squadron commander, and one chief of maintenance quality control.

The consensus opinion of these officers is that individual performance is important to the organization and is not limited to line technicians. The officers surveyed felt that although the rating forms designed for this study might be useful, they might not be valid if they were used as APRs. No one, however, was able to provide a more appropriate way to measure performance. One significant limitation to the gradations of quality used on the rating form was noted: many maintenance tasks require only compliance, with no gradations to the quality of work required.

All of the officers considered quality and quantity of performance to be important, although in some cases they felt that one cannot be considered independently of the other. All considered quality of performance to be absolutely overriding in importance compared to quantity of performance, the only exception to this being in times of critical emergencies, such as wartime.

This very small sample of opinions may not be representative of management viewpoints on the subject for the

entire Air Force. The author did feel, however, that some feedback from those who might be using this information would be useful. It was somewhat surprising to discover that quality of performance was considered to be much more important than quantity by all of the officers surveyed.

Summary

The results and analyses of this study have been presented in this chapter. Considerations included study of the sample to determine if the entire organization was well represented and a careful analysis of the sample data's statistical properties. In particular, the data was analyzed for applicability to regression analysis. The rating distributions were also tested for goodness-of-fit to normal distributions. Quantity and quality of performance ratings were also compared and an attempt was made to validate the quality ratings using MSEP results. Finally, maintenance officer opinions concerning quantity and quality of performance were summarized. All of these analyses will be discussed and interpreted in the next chapter.

Chapter 5

DISCUSSION

The purpose of this study is to determine the following:

1. What is the best research method for evaluating or measuring performance of aircraft maintenance technicians in the United States Air Force?
2. Does this method for evaluating or measuring performance provide useful and valid statistical data?

The discussion of the findings will therefore cover the performance appraisal method selected and the statistical evaluation of the appraisal method. This discussion may lead to findings that revise the existing body of knowledge concerning subjective performance appraisals or improve research methodologies.

The Performance Appraisal Method

The recommended performance appraisal method (Appendices B and C) was developed through a review of the literature on the subject. The literature review was necessary because existing rating schemes are either not applicable to statistical analysis, highly inflated, or unuseable for research. Airmen Performance Ratings (APRs) and civil service Merit Ratings are used for administrative

purposes of promotion and demotion and seldom reflect job performance alone. These ratings also tend to be inflated. Callander (1979:4) reports that airmen have average APR scores of 8.5 on a 9.0 scale, an inflated rating that is not useable for research. Proficiency ratings are either paper and pencil theory tests (Skill Knowledge Tests) or MSEP evaluations. For instance, only approximately 77% of the technicians at the test base, Williams AFB, had received MSEP appraisals due to the high turnover rate of personnel. Furthermore, SKT evaluations are not applicable to civil service maintenance technicians, who make up 17% of the line technicians at Williams AFB. Since existing performance data was not applicable, a new rating scheme was developed based on a review of the literature and the restrictions imposed by the maintenance organization.

The size of the Air Force maintenance organization required a measurement scheme applicable to civilian and military technicians of all races and sexes performing many tasks ranging from servicing aircraft to repairing missile guidance systems. It was thus apparent that the organization size and structure restricted useable performance measures to general criteria, such as quantity and quality of performance based on subjective appraisals by supervisors. The measurement of quantity and quality of performance presented a difficult problem.

Of the appraisal methods reviewed, the only applicable methods appeared to be straight ranking and the use of rating scales. Both methods are based on subjective appraisals by immediate supervisors. Barrett (1966:71) reported that the use of graphic scales following a forced ranking procedure increases the accuracy of the ratings. This method was adapted for this study and should have, in theory, provided rating scale performance values which were normally distributed.

The actual rating scale format was designed to minimize errors of leniency (see Appendices B and C). These recommended rating forms were appropriate based on a review of the literature concerning appraisals and on the nature of the Air Force maintenance organization. The suggested rating forms have face validity if previous research conclusions are accepted.

The following sections discuss the statistical qualities that resulted from an actual test of the rating forms within one Air Force aircraft maintenance organization. In addition, the opinions on the usefulness of the rating forms which were collected from several maintenance officers will be discussed.

Evaluation of the Rating Method

The evaluation of the rating method was conducted at Williams AFB, Arizona. This is a relatively small pilot training base utilizing jet aircraft which are mechanically simple

compared to other aircraft in the Air Force inventory--the T-37 and T-38 jet trainers. Thus, the conclusions of this study may be limited by the restriction of the test to one training base. One prerequisite of the rating form tested was its applicability to both military and civilian maintenance technicians. Both types of technicians were included in the sample drawn from the maintenance organization.

Test Sampling Procedures

The random sampling procedure resulted in representative proportions of civilian and military technicians. Rating of the technicians proved to be of no difficulty to any of the supervisors. It is significant to note that a majority of technicians within the FMS squadron were civilians or had attained the rank of sergeant or better. The OMS squadron, conversely, was primarily military with the majority being airmen. It can generally be concluded that the civilians and the Air Force personnel with sergeant rank and above have much more work experience than the airmen.

The sampling procedure also provided a representative proportion of technicians from each of the OMS branches. The FMS branches were not, however, proportionately represented by the sample. This uneven distribution was primarily due to the large number of independent sections or shops in FMS having very few people; the swing and mid shifts in the more heavily populated shops were excluded, as fewer than three technicians worked for a supervisor at any one time.

In general, the sampling procedure was extremely successful in providing a representative segment of the organization. By randomly selecting supervisory groups rather than individual technicians, the rating process was expedited. And by randomly selecting individuals from the supervisory groups, the researchers controlled who would be rated and avoided any bias had the supervisors selected the technicians themselves.

Rating Distributions

The skewed nature of the quantity and quality histogram distributions (Figures 14 and 15) was not expected. The rating forms were designed to produce symmetric, normal distributions that reflect the relative performance of technicians. There are two possible explanations for the consistently high, skewed quality ratings: (1) the supervisors as a whole may be extremely lenient or (2) the supervisors may be interested in consistent compliance and no more. There is no way to differentiate between these two explanations. It is possible, however, that many maintenance tasks are designed to be simple so that the technician, according to one survey response (Appendix K), "either can do the job, or he can't with no gradations in quality." If supervisors rate technicians in this light, then the quality histograms would be expected to display the skewed nature observed in the test data. It would be interesting to analyze technician tasks to discover if quality gradations

exist or if tasks are considered to be done or not done without quality gradations. At any rate, these quality of performance distributions were considered useful for lack of any other better performance measures.

The MSEP inspection distributions are more severely skewed than the research sample quality ratings (Figures 16 and 17) and omit some 20% of the line technicians. Sauer, Campbell, and Potter (1977) observed the same difficulty with MSEP data and also determined that the data was not applicable as a measure of performance in constructing mathematical models.

The quantity of performance ratings exhibited distributions which were less skewed than the quality ratings, and the quantity distributions were significantly different. The difference in the distributions was established using the Kolmogorov-Smirnov test (Appendix I). Despite the differences between quantity and quality, the quantity ratings were still not symmetric and still had mean values greater than the scale mid-point. The difference in the ratings is worthy of note, but the skewed distributions could also be the result of a halo effect from the quality ratings or, again, rater leniency. There is no way to differentiate between these influences. In fact, there is no existing data with which to compare quantity ratings. This information may give an added dimension to technician performance that is not currently considered but may be important in contingency situations.

Rating Distributions and Normality

The skewed nature of all the sample distributions makes it unlikely that they represent normal populations (Table 7). This does not create a serious problem for regression analysis but does restrict the types of tests that can be used for comparing distributions.

The skewed distributions should be used with caution in any regression analysis. It is possible that the error terms of the regression model may not have constant variance (heteroscedasticity) as a result of skewness. According to Neter and Wasserman (1974:123), this condition can best be investigated after a preliminary regression model is developed and, if necessary, a transformation of the performance data can be made to correct the problem. It is often the case that such a transformation also helps to normalize the data.

The skewed distributions do present problems with tests that compare distribution means or variances. In particular, the equality of variances is questionable in this situation. This means that non-parametric tests or tests which do not require equal variances should be used to compare distributions.

Quantity and Quality

The skewed distributions made it difficult to determine if any difference existed between quantity and

quality ratings. The histogram distributions were based on larger cell intervals than initially intended due to the grouping of ratings around scale numbers. This, in turn, produced large sample variances ($s^2=4.0$) and many ratings in separate distributions with the same values. As a result, parametric tests for comparing means had reduced power, while most non-parametric tests for comparing means were not useable due to the tied rating values.

To overcome these difficulties, conservative tests were used. The Kolmogorov-Smirnov non-parametric test showed that the quality and quantity rating distributions were different for both squadrons. The Behrens-Fisher approximate "t" test indicated that the means of the quantity and quality distributions for OMS and FMS differed. The quantity and quality distribution means for FMS were accepted as different with a Type I risk of 10% and a Type II risk of 13%. The quantity and quality distribution means for OMS were accepted as different with a Type I risk of 5% and a Type II risk of 30%. These differences indicate that quantity and quality ratings should be considered separately in evaluating technician performance. The differences do not indicate that technicians' ratings on one factor will not be reflected in the other factor.

The difficulties encountered in evaluating the ratings might be overcome by revising the scale format to eliminate the bunching of ratings around scale numbers and by thus attempting to force normal distributions. For now,

however, the data is useful in regression analysis with certain restrictions and the results do differentiate between quantity and quality ratings. Whether or not the ratings are valid when compared with existing data will be discussed in the next section.

Validity and Association

If the theoretical basis of the tests is sound and if the maintenance officers surveyed are to be believed, the performance ratings do have face validity. It is difficult, however, to find any agreement between the quality ratings and existing MSEP data.

No comparative data exists for the quantity ratings. It is interesting to note that supervisors tended to associate high quantity and high quality ratings for technicians. This could be due to a halo effect or could simply be based on the opinion that quantity and quality of performance are related.

Attempting to validate the quality ratings using MSEP data was not very successful. The limitations of the MSEP data were particularly difficult to deal with. More than 20% of the technicians received no inspections due to the rapid turnover of personnel and in spite of a local policy of administering an inspection every eighteen months. The resulting MSEP data, even after an attempt was made to interpret the raw data in relation to performance baselines, proved to be even more skewed and unuseable than the rating

distributions. It should be noted that the MSEP data covers two years and is utilized to determine section or branch trends and not to reflect individual performance. MSEP inspections also emphasize compliance and safety and not gradations of quality performance. The MSEP data was still used--in spite of these limitations--in an attempt to validate quality ratings, as it was the only existing record available. The correlation analysis revealed no significant relation between MSEP and the quality ratings.

When only technicians with two or more MSEP inspections were included in the correlation analysis, a low but significant non-zero (.368) correlation was noted for OMS, while no correlation at all existed for FMS. The revised OMS MSEP data in fact exhibited a normal distribution and a mean value somewhat lower than the OMS quality ratings; thus, if such data existed for all OMS personnel, it might be a superior measure of performance to the ratings. As for the FMS MSEP data, it may have remained highly skewed due to the difficulty of inspecting highly technical tasks, the relatively fewer average inspections, or the overall high experience level within the squadron. At any rate, the FMS quality ratings are superior to MSEP data for statistical research.

In summary, the quality and quantity ratings could not be conclusively validated using MSEP results. It does appear, however, that without more frequent MSEP inspections of all personnel, the performance rating distributions

provide more useable and representative data for use in statistical research. The performance ratings at least provide face validity.

Maintenance Officer Opinions

The limited number of maintenance officer opinions solicited concerning the quantity and quality of performance ratings produced some interesting results. Although quantity and quality were considered to be important considerations, they were not considered to be independent of one another. The majority of officers also felt that quality was the more important factor compared to quantity in all but the most dangerous national emergencies. This is surprising since the author has personally witnessed many officers pressuring technicians to do repairs rapidly. For this reason it is felt that both ratings should be of significant value to maintenance managers even if the emphasis continues to remain on quality performance.

Summary

The purposes of this study were to provide a method for evaluating or measuring the performance of Air Force maintenance technicians and to provide useful and valid performance data for statistical analysis. Based on a review of the literature, the performance evaluation method developed for this study provided a measure which (1) was understood by managers and supervisors, (2) was applicable

to both military and civilian technicians, (3) was applicable to different types of performance tasks, and (4) provided a performance measure through many levels of weapons systems maintenance. The statistical properties of the rating method, however, were discovered to be less than completely ideal.

The test sample revealed particular statistical properties which were not expected based on the literature review; particularly surprising were the skewed quantity and quality of performance distributions. These skewed distributions made the data applicable to regression analysis but with reservations concerning heteroscedasticity. They also made comparative tests difficult. Indeed, the nature of the histogram distributions made the use of most non-parametric tests impossible and reduced the power of parametric tests even with large sample sizes. Even given these considerations, however, it can be said that the data from the recommended performance rating method is useable for regression analysis and does differentiate between quality and quantity of performance.

These particular conclusions regarding the test statistics cannot be compared to the results of tests from other rating forms since such results do not appear to be widely reported. It was certainly surprising to discover that a method based upon previous theory and research did not produce the symmetrical distributions other authors reported; it should be mentioned that most reports did not

include data on tests for symmetry or equality of variances or means let alone the sample rating distributions.

Attempts to validate the quality of performance data were marginally successful for one squadron. The existing MSEP data used for the comparisons was even more skewed than the performance rating distributions and omitted some 20% of the sample. More work needs to be done to determine if the MSEP and the rating distributions were skewed as a result of an emphasis on technician task compliance or simply on rater and inspector leniency. At any rate, the data does not support the use of forced normal distributions for technician rankings on quantity and quality of performance.

It is evident that although the present method has limitations it is superior to existing information on individual technician performance. The potential exists for carefully monitoring the quality of maintenance technician performance using MSEP data, if Air Force management feels it useful. This information does not exist now, however, while no information is even available concerning the quantity of performance. Thus, for lack of any superior system, the subjective ratings of technician quantity and quality of performance are useful and acceptable as sources of performance statistics.

Chapter 6

CONCLUSION

One of the greatest needs of managers of the military weapons system maintenance complex is to measure accurately how well individuals perform on the job. Individual job performance is one of the bases for performance by the entire organization. If the effectiveness of weapons system maintenance is to be improved, then individual performance must be measurable and subject to improvement.

Quantifying job effectiveness is, however, difficult. Decades of research by psychologists and personnel experts have failed to provide definitive answers to the question of how to measure performance or effectiveness. The main purpose of this study was to find or develop some method for evaluating and measuring the performance of aircraft maintenance technicians in the United States Air Force. This evaluation method, once developed, is to be used as a performance measure of manpower effectiveness in another research effort. The purpose of this subsequent research effort will be to develop a model or models for predicting or evaluating the effectiveness of maintenance technician performance (see Young, 1978:15).

The recommended performance appraisal method (Appendices B and C) was developed through a review of the literature on the subject. The literature review was

necessary as existing appraisal methods such as APRs, MSEP, or SATs either were not applicable to statistical analysis, were highly inflated, or provided incomplete and non-current coverage of the organization. The method which was developed relied on subjective supervisor appraisals of a technician's quantity and quality of performance. This suggested method has face validity, if previous research conclusions are to be accepted.

The evaluation of the performance appraisal method was conducted within the aircraft maintenance organization of one pilot training Air Force Base, Williams AFB, Arizona. The evaluation at one base limits the generality of the test conclusions. A sample selection method was developed that actually paired supervisors and their subordinate technicians and provided a representative portion of the maintenance organization. A sample size of 20% of the organization provided adequate statistical test errors, with the exception that the rating scale and resulting rating distribution increased the test error. A change in the rating scale eliminating numbered gradations in quality and quantity might eliminate this problem.

The test evaluation of the performance appraisal method also resulted in skewed quantity and quality ratings, results which were not expected based on the literature review. It is difficult to determine from this present study if these skewed ratings represented an emphasis on technician task compliance or simply on rater leniency.

The ratings were certainly more complete and less skewed than existing Maintenance Standardization and Evaluation (MSEP) personnel inspections. The skewed ratings do introduce some restrictions in the development of a regression model to predict or evaluate the effectiveness of maintenance technician performance; special care should be taken to identify and correct for heteroscedasticity. Quality and quantity ratings were differentiable and should both be used to represent performance, although the maintenance officers surveyed indicated that quality of performance is more essential to mission accomplishment than quantity of performance.

Attempts to validate the quality of performance were marginally successful for the Operational Maintenance Squadron (OMS) involved in the test. The attempts to validate quality ratings for the Field Maintenance Squadron (FMS) were unsuccessful. MSEP personnel inspection data was used for these comparisons and proved to be highly inflated and non-representative of the organization. No data existed with which to compare the quantity of performance ratings.

Despite these difficulties, the performance rating method provides useful data with face validity which can be obtained for a representative segment of an Air Force maintenance organization. It is evident that the rating data must be used with care in attempting to develop a model of organizational effectiveness.

Contributions and Future Considerations

This study makes several contributions to the field of performance appraisal within Air Force organizations. The recommended performance rating method is new and provides useful information. In testing the performance rating method, a sample selection technique was developed that provided input from supervisors and their immediate subordinates and provided a representative segment of the maintenance organization. Many previous studies failed to ensure that supervisors were actually evaluating their subordinates. Most significantly, an analysis of the test results provided information on the statistical effects of skewed distributions, of rating scales based on numbered gradations of performance, and on the use of histograms in situations where non-parametric statistical tests are required. These contributions were offset, however, by areas which were found to require further evaluation.

A superior rating scale might be suggested from the results of this study. Such a scale would have only end- and mid-point descriptions (e.g., "slowest," "where most perform," and "fastest"), no numbered gradations, and one scale mark at the mid-point. Values for such a scale would be recorded by the researcher, who would be using a separate numbered scale. Such a scale should provide symmetrical performance distributions with small variance about the means and correspondingly low Type II errors in comparative

statistical tests. Such a scale, however, would have low face validity and might not be acceptable to supervisors or maintenance managers.

Whatever rating method is used, it should certainly be tested at more locations than the one base used in this study. It might also be wise to include second-level supervisors' ratings as a comparative and controlling influence on the ratings of technician performance by immediate supervisors.

It is unfortunate that existing performance appraisal information cannot be used to evaluate technician performance for this research effort. Existing data would be superior to the proposed performance ratings in terms of validity and acceptability. However, Airmen Performance Ratings (APRs) and Skill Knowledge Tests (SKTs) are not appropriate. The MSEP data, on the other hand, could be quite useful if Air Force requirements (AFM 66-1) were more specific regarding regular inspections for individuals or provided a method for randomly selecting inspections samples. As it is used now, MSEP is supposed to provide trend analysis data for maintenance sections or branches, but this data is based on invalid sampling procedures. The MSEP data was suspiciously skewed with many perfect appraisals and, as a result, was not strongly related to current supervisor ratings. MSEP data could, however, provide more useful information if Air Force personnel inspection criteria were revised.

Final Comment

It is evident from this study that although the recommended subjective performance appraisal method has limitations, it is superior to existing information on individual technician performance. The appraisal method has face validity for the evaluation of aircraft maintenance technicians in the United States Air Force. It also provides useful statistical data. Existing information on the quality of technician performance is potentially useful, but it falls short of being representative and inclusive. Little information even exists concerning the quantity of technician performance. Thus, for lack of any superior system, the subjective ratings of technician quantity and quality of performance developed in this study are useful and acceptable sources of performance statistics.

REFERENCES

- Barrett, Richard S. Performance Rating. Chicago: Science Research Associates, 1966.
- Bayroff, A.G., H.R. Haggerty, and E.A. Rundquist. "Validity of Ratings as Related to Rating Techniques and Conditions." Personnel Psychology, 1954, 7, 93-112.
- Bendig, A.W. "Reliability and the Number of Rating Scale Categories." Journal of Applied Psychology, 1954, 38, 38-40.
- Berk, Kenneth N., and Ivor S. Francis. "A Review of the Manuals for BMDP and SPSS." Journal of the American Statistical Association, 1978, 301, 65-71.
- Beyer, William H., ed. CRC Standard Mathematical Tables. West Palm Beach, Fla.: CRC Press, 1978.
- Bittner, R. "Developing an Industrial Merit Rating Procedure." Personnel Psychology, 1948, 1, 403-432.
- Bradley, James V. Distribution-Free Statistical Tests. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1968.
- Brozden, H.E., and E.K. Taylor. "The Theory and Classification of Criterion Bias." Educational and Psychological Measurement, 1950, 10, 159-186.
- Callander, Bruce. "Hikes: 1 out of 3 Makes E-7." Air Force Times, July 2, 1979, p.4.
- Campbell, J.P., and others. Managerial Behavior, Performance and Effectiveness. New York: McGraw-Hill, 1970.
- Campbell, J.T., E.P. Prien, and L.G. Brailey. "Predicting Performance Evaluations." Personnel Psychology, 1960, 13, 435-440.
- Cummings, L.L., and D.P. Schwab. Performance in Organizations. Glenview, Ill.: Scott, Foresman, Inc., 1973.
- Davies, Owen L. The Design and Analysis of Industrial Experiments. London: Oliver and Boyd, 1956.
- Dixon, W.J., ed. BMDP Biomedical Computer Program. Los Angeles: University of California Press, 1977.

- Dudewicz, Edward J. Introduction to Statistics and Probability. New York: Holt, Rinehart, and Winston, 1976.
- Dunnette, M.D. "A Note on the Criterion." Journal of Applied Psychology, 1963, 41, 251-254.
- Flanagan, J.C. "The Critical Incident Technique." Psychological Bulletin, 1954, 51, 327-358.
- Foley, John P. Evaluating Maintenance Performance: An Analysis. AFHRL-TR-75-57(1), AD-053-475. Wright-Patterson AFB, OH: Advanced Systems Division, Air Force Human Resources Laboratory, December 1974.
- Freund, John E., and Irwin Miller. Probability and Statistics for Engineers. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1977.
- Gibbons, Jean D., Non-parametric Methods for Quantitative Analysis. New York: Holt, Rinehart, and Winston, 1970.
- Glaser, R., and A.J. Nitko. "Measurement in Learning and Instruction." In R.L. Thorndike (Ed.), Educational Measurement. Washington: American Council on Education, 1971, 525-570.
- Guilford, Joy P. Psychometric Methods. New York: McGraw-Hill, 1954.
- Guion, Robert M. Personnel Testing. New York: McGraw-Hill, 1965.
- Habbe, S. "Marks of a Good Worker." Management Record, 1956, 13, 168-170.
- Hollander, Myles, and Douglas A. Wolfe. Nonparametric Statistical Methods. New York: John Wiley and Sons, 1973.
- Hollingworth, H.L. Judging Human Character. New York: Appleton, 1922.
- Holly, W.H., H.S. Field, and N.J. Barnett. "Analyzing Performance Appraisal Systems: An Empirical Study." Personnel Journal, 1976, 55, 457-459.
- Jones, L.V., and L.L. Thurstone. "The Psychophysics of Semantics: An Experimental Investigation." Journal of Applied Psychology, 1955, 30, 31-36.
- Larin, Lawrence L. Statistics for Modern Business Decisions. New York: Harcourt, Brace, Jovanovich, Inc., 1973.

- Lawler, E.E., III. "The Multitrait-Multirater Approach to Measuring Managerial Job Performance." Journal of Applied Psychology, 1967, 51, 369-381.
- Lewin, Arie Y., and A. Zwany. "A Model Literature Critique and a Paradigm for Research." Personnel Psychology, 1976, 29, 423-447.
- Locker, Alan H., and K.S. Teel. "Performance Appraisal--A Survey of Current Practices." Personnel Journal, 1977, 56, 245-247.
- Lopez, Felix M. Evaluating Employee Performance. Chicago: Public Personnel Association, 1968.
- McDonnell, James A. "Distaff Mechanics Doing OK." Air Force Magazine, 1979, 62, 78-79.
- McGregor, Douglas. "An Uneasy Look at Performance Appraisal." Harvard Business Review, 1967, 35, 80-94.
- Meister, D., D.L. Finley, and E.A. Thompson. Relationship Between System Design, Technician Training, and Maintenance Job Performance on Two Autopilot Subsystems. AFHRL-TR-70-20. Bunker-Ramo Corporation, Wright-Patterson AFB, OH: Advanced Systems Division, Air Force Human Resources Laboratory, 1971.
- Millard, Cheedle W., F. Luthans, and R.L. Otteman. "A New Breakthrough for Performance Appraisal." Business Horizons, 1976, 19, 66-73.
- Miner, J. "Management by Appraisal: A Capsule Review and Current References." Business Horizons, 1968, 11, 33-94.
- Muller, Mervin E. "A Review of the Manuals for BMDP and SPSS." Journal of the American Statistical Association, 1978, 361, 71-80.
- Neter, John, William Wasserman, and G.A. Whitmore. Fundamental Statistics. Boston: Allyn and Bacon, Inc., 1973.
- Neter, John, and William Wasserman. Applied Linear Statistical Models. Homewood, Ill.: Richard D. Irwin, Inc., 1974.
- Obradovic, J. "Modification of the Forced Choice Method as a Criterion of Job Proficiency." Journal of Applied Psychology, 1970, 54, 228-233.

- Ronan, W.W., C.L. Anderson, and T.L. Talbert. "A Psychometric Approach to Job Performance: Fire Fighters." Public Personnel Management, 1976, 5, 409-422.
- Rush, C.H. "A Factorial Study of Sales Criteria." Personnel Psychology, 1953, 6, 140-157.
- Sauer, D.W., W.B. Campbell, and N.R. Potter. Human Resource Factors and Performance Relationships in Nuclear Missile Handling Tasks. AFHRL-TR-76-85, AFWL-TR-76-30. Wright-Patterson AFB, OH: Advanced Systems Division, Air Force Human Resources Laboratory, 1977.
- Snedecor, George W., and William G. Cochran. Statistical Methods. Ames, Iowa: The Iowa State College Press, 1956.
- Stevens, S.N., and E.F. Wonderlic. "An Effective Revision of the Rating Technique." Personnel Journal, 1934, 13, 125-134.
- Swezey, R.W., and R.B. Pearlstein. "Developing Criterion-Referenced Tests." JSAS Catalog of Selected Documents in Psychology, 1975, 5, 227.
- Taylor, E.K., R.S. Barrett, J.W. Parker, and L. Martens. "Rating Scale Content: II. Effect of Rating on Individual Scales." Personnel Psychology, 1958, 11, 512-533.
- Taylor, E.K., and Grace E. Manson. "Supervised Ratings; Making Graphic Scales Work." Personnel, 1951, 27, 504-514.
- Thornton, G. "The Relationship Between Supervisory and Self-Appraisal of Executive Performance." Personnel Psychology, 1968, 21, 441-455.
- Travers, R.W. "A Critical Review of the Validity and Rationale of the Forced Choice Technique." Psychological Bulletin, 1951, 48, 62-70.
- Uhrbrock, R.S. "2000 Scaled Items." Personnel Psychology, 1961, 14, 375-420.
- U.S. Air Force. Air Force Manual (AFM) 66-1, Maintenance Management. 10 vols. Washington: Government Printing Office, 1975 and 1977.
- Vanzelst, R., and W. Kerr. "Worker Attitude Toward Merit Rating." Personnel Psychology, 1953, 6, 159-172.

Whitla, Dean K., and John E. Tirrell. "The Validity of Ratings of Several Levels of Supervisors." Personnel Psychology, 1953, 6, 461-466.

Wikstrom, W.S. Managing by and with Objectives. National Industrial Conference Board, Personal Study No. 212, 1968.

Wiley, L.N. Task Level Performance Criteria Development. AFHRL-TR-77-75, AD-055-694. Brooks AFB, TX: Occupational and Manpower Research Division, Air Force Human Resources Laboratory, December 1978.

Young, H.H. Performance Effectiveness in the Air Force Maintenance System: Preliminary Report and Design Report. Wright-Patterson AFB, OH: Advanced Systems Division, Air Force Human Resources Laboratory, 1978. (Internal Report)

APPENDIX A

MAINTENANCE TECHNICIAN SURVEY
PRIVACY STATEMENT

The attached survey is part of a research effort being conducted by Arizona State University under contract with the Air Force Office of Scientific Research, and with the cooperation of the Air Force Human Resources Laboratory, Advanced Systems Division, WPAFB, Ohio. The purpose of the survey is to further identify factors which influence performance effectiveness in maintaining Air Force aircraft and missile systems.

Your participation in the Survey is voluntary but strongly desired. Your responses will be held confidential and in no way will impact upon your career nor upon the squadron to which you are assigned. Headquarters USAF Survey Control Number 80-11 has been assigned to this survey.

A. Authority:

- (1) 5 U.S.C. 301, Departmental Regulations; and
- (2) 10 U.S.C. 8012, Secretary of the Air Force, Powers, Duties, Delegation by Compensation;
- (3) DOD Instruction 1100.13, 17 Apr 68, Surveys of Department of Defense Personnel; and/or
- (4) AFR 30-23, 23 Sep 76, Air Force Personnel Survey Program.

B. Principal Purposes: To collect information from Air Force and civilian squadron maintenance personnel concerning their perceptions of factors which influence their performance effectiveness. To initiate the development of an Air Force Maintenance Performance Effectiveness Model based on the survey results and other inputs.

C. Routine Uses: Data will be used for research purposes in initiating a predictive model of maintenance performance effectiveness.

D. Participation is voluntary. However, your cooperation is requested.

E. No adverse action of any kind may be taken against any individual who elects not to participate in any or all of this survey.

QUALITY OF PERFORMANCE RANKINGS

Highest
10

[illegible]

APPENDIX D

DATA

OMS Tech.	Quant.	Qual.	MSEP	Raw MSEP
1	8.0	6.0	.667	1/3,1/3
2	8.0	9.0	.888	0/2,0/3,1/3
3	8.0	10.0	.611	0/3,1/3,1/3,1/3,3/3,1/3
4	7.0	8.0		
5	8.0	6.0	.563	2/2,7/12,0/3,2/12
6	1.0	2.0		
7	5.0	8.0	.333	12/12,1/3
8	6.0	7.0	.333	2/3,2/3
9	5.0	8.0	.833	0/3,1/3
10	7.0	9.0	.667	0/3,2/3
11	9.0	9.0	.667	0/3,2/3
12	8.0	8.0	.888	0/3,0/3,1/3
13	8.0	8.0	.528	5/5,5/12,0/3
14	6.0	8.0	.611	0/3,0/3,0/3,2/3,2/3
15	6.0	6.0	.417	1/3,0/3,3/3,3/3
16	9.0	9.0	.500	6/6,0/3
17	2.0	4.0	.536	0/3,7/12,0/3,0/3,2/3, 3/3,3/3
18	8.0	8.0	.000	3/3
19	6.0	6.0	.777	1/3,1/3,0/3
20	8.0	8.0	.444	0/3,2/3,9/9
21	7.0	7.0	.381	0/3,1/3,0/3,0/3,3/3, 9/9,1/3
22	2.0	3.0	.667	1/3,1/3
23	4.0	4.0		
24	9.0	9.0	.733	0/3,1/3,2/3,1/3
25	8.0	9.0		
26	7.0	7.0	.263	6/11,2/3,3/3
27	9.0	9.0	.625	3/4,0/3
28	9.0	9.0	.750	2/3,0/3,0/3,3/9
29	7.0	5.0	.472	0/3,3/3,1/3,7/9
30	8.0	9.0	.333	0/3,0/3,1/3,2/3
31	9.0	9.0	.778	0/3,1/3,1/3
32	5.0	3.0	.000	1/1,3/3
33	7.0	8.0		
34	7.0	7.0	.200	3/5,2/2
35	6.0	6.0	1.00	0/3
36	5.0	8.0	1.00	0/3
37	5.0	7.0	1.00	0/3
38	6.0	6.0		

APPENDIX D (CONT.)

DATA				
OMS Tech.	Quant.	Qual.	MSEP	Raw MSEP
39	4.0	3.0		
40	7.0	9.0	.722	5/9,0/3
41	3.0	2.0		
42	10.0	7.0	1.00	0/3,0/3,0/3
43	9.0	8.0	1.00	0/3
44	8.0	9.0	.667	0/3,0/3,3/3,1/3
45	7.0	8.0		
46	9.0	9.0	1.00	0/3,0/3
47	2.0	3.0	.689	3/5,0/3,1/3
48	8.0	7.0	.555	3/3,0/3,3/9
49	5.0	5.0		
50	1.0	4.0	1.00	0/3
51	6.0	7.0	1.00	0/3
52	5.0	6.0		
53	10.0	10.0	.861	0/1,0/1,5/12
54	9.0	10.0	1.00	0/3,0/3
55	7.0	9.0		
56	8.0	9.0		
57	8.0	9.0		
58	7.0	8.0		
59	9.0	9.0	.833	0/3,1/3
60	7.0	8.0	.667	1/3,1/3
61	7.0	8.0	.773	5/11,0/3
62	5.0	7.0	.733	0/3,2/2,0/3,0/3,3/9
63	7.0	8.0	1.00	0/3
64	1.0	3.0	.500	1/1,0/3
65	3.0	6.0	.818	0/3,6/11,0/3
66	7.0	8.0	.583	0/3,12/12,0/3,2/3
67	6.0	6.0	.667	1/3,1/3
68	7.0	9.0	.333	2/3,2/3
69	7.0	7.0	1.00	0/3
70	8.0	9.0		
71	6.0	7.0	1.00	0/3,0/3
72	8.0	8.0	1.00	0/3,0/3,0/3,0/3
73	6.0	8.0	1.00	0/3,0/3
74	7.0	9.0	1.00	0/2,0/3,0/3,0/3,0/3
75	8.0	8.0	.542	7/12,1/3
76	4.0	6.0		
77	8.0	8.0	.778	1/3,0/3,1/3

APPENDIX D (CONT.)

DATA

OMS Tech.	Quant.	Qual.	MSEP	Raw MSEP
78	8.0	9.0	1.00	0/3,0/3,0/3
79	5.0	6.0	1.00	0/3
80	6.0	6.0	1.00	0/3
FMS				
1	9.0	9.0	.833	0/3,1/3
2	9.0	9.0	1.00	0/3,0/3
3	6.0	9.0	.778	0/3,0/3,2/3
4	8.0	9.0	1.00	0/3
5	8.0	9.0	1.00	0/3,0/3
6	1.0	2.0	.833	1/3,0/3
7	10.0	10.0	1.00	0/3
8	9.0	9.0	1.00	0/3
9	1.0	1.0		
10	6.0	5.0	1.00	0/3,0/3,0/3
11	8.0	10.0	.833	0/3,1/3
12	3.0	3.0	.166	2/3,3/3
13	6.0	9.0	1.00	0/2
14	5.0	7.0		
15	2.0	4.0		
16	3.0	3.0		
17	5.0	7.0	.500	1/2,1/2
18	5.0	6.0		
19	5.0	8.0	1.00	0/3
20	6.0	8.0	.267	3/3,2/3,3/3,1/3,2/3
21	8.0	9.0		
22	7.0	8.0	1.00	0/3,0/3
23	7.0	8.0	1.00	0/3
24	7.0	8.0	.667	0/3,0/3,3/3
25	8.0	9.0	1.00	0/3,0/3
26	8.0	8.0	.888	0/3,0/3,1/3
27	6.0	6.0		
28	6.0	7.0		
29	1.0	1.0		
30	8.0	8.0	1.00	0/3
31	2.0	2.0		
32	6.0	7.0		
33	10.0	9.0		
34	8.0	9.0	.888	1/3,0/3,0/3
35	6.0	8.0	1.00	0/3

APPENDIX D (CONT.)

DATA				
FMS Tech.	Quant.	Qual.	MSEP	Raw MSEP
36	6.0	7.0		
37	9.0	9.0	1.00	0/3,0/3
38	7.0	8.0	1.00	0/3
39	9.0	9.0	1.00	0/3,0/3
40	9.0	8.0	1.00	0/3,0/2
41	8.0	7.0	.833	0/3,1/3
42	6.0	4.0	1.00	0/3
43	5.0	6.0	1.00	0/3,0/3
44	4.0	3.0		
45	6.0	5.0	.833	1/3,0/2
46	7.0	7.0	1.00	0/2
47	8.0	10.0	.750	0/3,0/3,1/3,2/3
48	6.0	6.0	1.00	0/3
49	6.0	10.0	1.00	0/2
50	4.0	7.0	1.00	0/3,0/3,0/3
51	9.0	8.0	.833	0/3,1/3
52	8.0	9.0	.500	0/3,2/2
53	5.0	6.0		
54	6.0	8.0	1.00	0/3
55	7.0	8.0	1.00	0/3
56	9.0	9.0	1.00	0/3
57	6.0	6.0	1.00	0/3
58	5.0	6.0	1.00	0/3
59	6.0	7.0	1.00	0/3,0/3
60	8.0	9.0	.833	1/3,0/3
61	9.0	9.0	.867	0/3,0/3,0/3,1/3,1/3
62	7.0	7.0	1.00	0/3
63	9.0	10.0	1.00	0/2,0/1
64	7.0	7.0		
65	9.0	9.0	1.00	0/2
66	7.0	8.0	1.00	0/3
67	5.0	7.0	1.00	0/3
68	6.0	8.0	.833	0/3,1/3
69	9.0	8.0	1.00	0/3
70	6.0	5.0	1.00	0/3
71	8.0	8.0	.833	0/3,1/3
72	8.0	9.0	1.00	0/3
73	7.0	6.0	1.00	0/3
74	7.0	7.0	1.00	0/3
75	9.0	10.0	1.00	0/3,0/3

APPENDIX D (CONT.)

DATA

FMS Tech.	Quant.	Qual.	MSEP	Raw MSEP
76	8.0	8.0	1.00	0/3
77	4.0	5.0		
78	6.0	7.0	1.00	0/3
79	9.0	9.0	1.00	0/3,0/3
80	9.0	8.0	1.00	0/3
81	3.0	5.0		
82	8.0	9.0	1.00	0/3
83	8.0	7.0	1.00	0/3
84	5.0	6.0		
85	7.0	8.0	1.00	0/3
86	5.0	9.0	1.00	0/3,0/3
87	8.0	7.0	1.00	0/3,0/3,0/3
88	7.0	8.0		
89	6.0	9.0	1.00	0/3

- Notes: 1. The Technician numbers that appear here are not the same as the code numbers used in the actual experiment. The numbers were changed to protect the identities of the technicians.
2. The Quantity and Quality ratings are based on a 10.0 scale.
3. The Raw MSEP data reflects the number of discrepancies found by inspectors versus the failure baseline for the particular task inspected.
4. The MSEP data is calculated by averaging the Raw MSEP data and subtracting this value from one.

Variable	Number	Mean	Median	Mode	Minimum	Maximum	Range	Variance	St. Dev.	Count	Percentage	Sum
Number of Distinct Values	11	1.000000	1.000000	1.000000	1.000000	1.000000	0.000000	0.000000	0.000000	11	100.00	11.00
Number of Values Counted	70	1.000000	1.000000	1.000000	1.000000	1.000000	0.000000	0.000000	0.000000	70	100.00	70.00
Number of Values Not Counted	19	1.000000	1.000000	1.000000	1.000000	1.000000	0.000000	0.000000	0.000000	19	100.00	19.00
Location Estimates		0.919714	0.923191	0.923191	0.919714	0.923191	0.000000	0.000000	0.000000	19	100.00	17.27
Some NPM Location Estimates		1.000000	1.000000	1.000000	1.000000	1.000000	0.000000	0.000000	0.000000	19	100.00	19.00
Wampel		1.000000	1.000000	1.000000	1.000000	1.000000	0.000000	0.000000	0.000000	19	100.00	19.00
Amurals		1.000000	1.000000	1.000000	1.000000	1.000000	0.000000	0.000000	0.000000	19	100.00	19.00
Turev		1.000000	1.000000	1.000000	1.000000	1.000000	0.000000	0.000000	0.000000	19	100.00	19.00

APPENDIX E (CONT.)

Statistical Data
FMS MSEP from BMDP (Dixon:1977) Detailed Data Description (P2D) program.

APPENDIX F

Chi-Square Goodness-of-Fit Calculation^a FMS Quality of Performance

1. H_0 : Sample is from Population $N(7.3, 2.1)$
- H_1 : Sample is from population having some other distribution.
2. v =number of classes-3 (estimate μ, σ , and Σe)^b=7-3=4
3. Type I risk, $\alpha=.001$.
4. Critical Region: Reject null hypothesis if $\chi^2 > (\chi^2_{\alpha=.001}=18.465)$
5. Calculation of Chi-Square ($n=89, \bar{x}=7.3; s=2.1$)

Quality Rating (x=Upper Class lim.)	Observed Frequency (f_o)	Normal Deviate ($z = \frac{x-7.3}{2.1}$)	Area to Left of x (.0918)	Area of Class Interval (.0918)	Expected Frequency (f_e)	Observed-Expected Freq. ($f_o - f_e$)	$(f_o - f_e)^2$ + .83	$(f_o - f_e)^2$ f _e
4.5-5.5	9	-1.33	.0918	.0918	8.17	-4.18	.6889	.08
5.5-6.5	5	-.86	.1949	.1031	9.18	-4.18	17.4724	1.90
6.5-7.5	16	+.10	.5398	.1571	13.98	-4.98	24.8004	1.77
7.5-8.5	22	+.57	.7157	.1878	16.71	-.71	.5041	.03
8.5-9.5	22	+1.05	.8531	.1759	15.66	+6.34	40.1956	2.57
>9.5	6	infin.	1.0000	.1374	12.23	+9.77	95.4529	7.81
	89		1.0000	.1469	13.07	-7.07	49.9849	3.82
				1.0000	89.00	0.00		$\chi^2=17.98$

6. Conclusion: Since $(\chi^2=17.98) < (\chi^2_{.001}=18.465)$, if the quality rating distribution were $N(7.3, 2.1)$ then there is a .1% chance that a sample would be obtained from the normal distribution with $\chi^2 > 18.465$.

^aCalculation and table sources: Freund (1977) and Lapin (1973).
^bOne degree of freedom is lost as the f_e 's must sum to n.

APPENDIX G

Effect of Skewness and Kurtosis
on Type I Error (CA)
(γ_1 =Skewness; γ_2 =kurtosis)

t-test Comparison of means of 2 groups ^a ($n^b=10$; $\alpha=.050$ or 5.0% ; $\delta_1=1$)		F-test Comparison of variances for 2 groups ^c (D.F. $c=4, 20$; $\alpha=.05$ or 5.0% ; $\gamma_1=1$)	
δ_2	Actual Risk (α)	δ_2	Actual Risk (α)
0.0	5.02%	0.0	4.49%
0.5	4.96%	0.5	5.54%
1.0	4.90%	1.0	6.58%

^aSource: Davies (1956:53-55); interpolation was used to find the actual risk for small kurtosis values.

^bFor t-test comparison of means, increasing n reduces actual error to α .

^cFor F-test comparison of variances, increasing n does not reduce the actual error to α .

APPENDIX H

F-Test for Equivalence of Sample Variances
FMS Quantity and Quality of Performance^a

1. H_0 : $\text{variance}_1 = \text{variance}_2$
 H_1 : $\text{variance}_1 \neq \text{variance}_2$
2. ν = degrees of freedom = $(n_1 - 1; n_2 - 1) = (88, 88)$
3. Risk = .05 = alpha
4. Critical Region: Accept H_0 if

$$\begin{array}{rcccl} F(\alpha/2; n_1 - 1, n_2 - 1) & \leq & F^* & \leq & F(1 - \alpha/2; n_1 - 1, n_2 - 1) \\ F(.025; 88, 88) & \leq & F^* & \leq & F(.975; 88, 88) \\ .645 & \leq & F^* & \leq & 1.55 \end{array}$$
5. Calculation of F:

$$F^* = \frac{s_1^2}{s_2^2} = \frac{4.2497}{4.2641} = .9966$$
6. Conclusion: cannot reject H_0 : $\text{variance}_1 = \text{variance}_2$

^aCalculation and table source: Neter and Wasserman(1974)

APPENDIX I

Comparison of Sample Distributions
FMS Quantity versus Quality of Performance

1. Kolmogorov-Smirnov non-parametric two sample test.^{a, b}

$H_0: F_1(x) = F_2(x)$ for all x .

$H_1: F_1(x) < F_2(x)$ for some x .

2. Risk: $P = .05^c$

3. Critical Region:

Reject H_0 if $D_{-.05} \geq [1.22 \sqrt{\frac{175}{89}}] = 1.22(.1499) = .1829$

4. Calculation of D_- :

x	$G_s(x)$	$G_q(x)$	$S_s(x)$	$S_q(x)$	$ S_s(x) - S_q(x) $
1	3	2	.03	.02	.01
2	5	4	.06	.05	.01
3	8	7	.09	.08	.01
4	11	9	.12	.10	.02
5	21	14	.24	.16	.08
6	41	23	.46	.26	.20*
7	54	30	.61	.44	.17
8	72	61	.81	.69	.12
9	87	83	.98	.93	.05
10	89	89	1.00	1.00	.00

$D_- = .20$

5. Conclusion: as $D_- \geq D_{-.05}$, reject the null hypothesis, conclude that $F_1 < F_2$.

^aSource: Gibbons (1976;252).

^bAssume that the quantity and quality ratings for each technician by the same supervisor are independent samples from two populations.

^cThe risk (P) value is based on the assumption that quantity and quality of performance are continuous variables. If these are continuous variables, then the P value should be considered conservative. (Gibbons: page 258).

APPENDIX I (CONT.)

Comparison of Sample Distributions
FMS Quantity versus Quality of Performance

1. Behrens-Fisher approximate "t" test, (Dudewicz;1976:309)
for the case where variance₁ ≠ variance₂.

$$H_0: \text{mean}_1 = \text{mean}_2$$

$$H_1: \text{mean}_1 \neq \text{mean}_2$$

- 2.

$$\nu = \text{degrees of freedom}^a = \frac{2(s_1^2/n_1 + s_2^2/n_2)^2}{[(s_1^2/n_1)/(n_1-1) + (s_2^2/n_2)/(n_2-1)]} \\ = 175.75 \text{ or } 175$$

3. Risk = α .05

4. Critical Region:

$$\text{Reject } H_0 \text{ if } |t^*| \geq t(\alpha/2; \nu)$$

5. Calculation of "t":

$$t^* = \frac{(x_1 - x_2) - (\mu_1 - \mu_2)}{(s_1^2/n_1 + s_2^2/n_2)^{1/2}} = 2.179$$

6. Conclusion: Reject H_0 : mean₁ ≠ mean₂

7. Type II error: Calculated for a given interval difference assuming equal variance (Freund;1977:217).

$$d = \frac{|\delta - \delta'|}{\sqrt{\sigma_1^2 + \sigma_2^2}} = .231$$

Using the tables for two-tailed tests ($\alpha = .05$)
from Freund (1977:497), the Type II error (β) equals .30.

^aWelch's formulation for degrees of freedom for the
Behrens-Fisher problem (Dudewicz;1976:311).

APPENDIX J

Correlation Analysis
for Quantity/Quality Associations and
Quality/MSEP Associations^a

1. All linear correlation coefficient calculations were made using the BMDP Bivariate Plot (P6D) program (Dixon; 1977). The bivariate plots and correlation coefficients are presented on pages 128 and 129. The coefficients were tested in the following manner:

$$H_0: \rho = 0$$

$$H_1: \rho \neq 0$$

2. Risk = $\alpha = .01$

3. Critical Region: Reject H_0 if

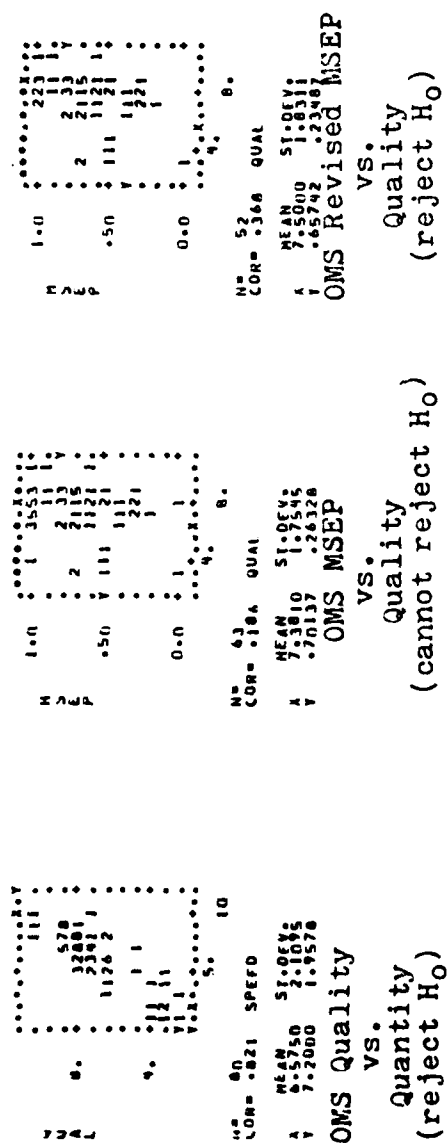
$$r \geq [r^*(n-2)] \quad (\text{from Snedecor; 1956:173})$$

4. Calculation:

$$r = \frac{\sum_j (x_{ij} - \bar{x}_i) (x_{kj} - \bar{x}_k)}{[\sum_j (x_{ij} - \bar{x}_i)^2 \sum_j (x_{kj} - \bar{x}_k)^2]^{1/2}}$$

5. Conclusions: See pages 128 and 129.

^aRevised MSEP data includes only personnel with two or more inspections.



APPENDIX J (CONT.)

Correlation Analysis for
Quantity/Quality Associations and
Quality/MSEP Associations

From BMDP Bivariate Plot (P6D) program

APPENDIX K MAINTENANCE OFFICER SURVEY

From: Capt. Joel R. Hickman

San Diego, CA, 11/11/71

To:

Dear Sir,

I am currently an AFIT student at Arizona State University, an assignment less demanding than my last two years as a flight safety officer at McChes. At present I am contributing to an Air Force sponsored study of maintenance performance, and plan to write my thesis on that subject. In January, upon completion of the thesis, the Air Force is sending my family and me to A.F. Sawyer AFB in Mississippi for a three year tour in Civil Engineering.

I am soliciting your opinion as a maintenance officer concerning the enclosed forms. These forms will be completed by supervisors and will be coordinated in our studies at AFIT and AFB. I would appreciate your comments on performance and on the appraisal forms in particular using the enclosed answer sheet.

Please return the answer sheet to me by mail, using the enclosed self-addressed envelope. Your identity will not be divulged although the study may be published. I would greatly appreciate your interest and assistance.

Joel R. Hickman

Joel R. Hickman, Capt, USAF

1. Name of Officer
2. Position
3. Performance
4. Quantity of Performance

APPENDIX K (CONT.)
ANSWER SHEET

1. Is individual performance important to the maintenance organization?
2. Are the ranking forms (Attachments 2 and 3) appropriate for appraising performance or can you suggest a better approach?
3. Are quantity and quality useful measures of performance?
4. Which do you consider to be more important, quantity or quality?
5. If one is more important than the other, can you indicate how much more important it is?

APPENDIX L

SURVEY ANSWERS

1. Is individual performance important to the maintenance organization?

Answer 1: Definitely!

Answer 2: Of course. Maintenance production is the sum of individual performances.

Answer 3: Without doubt. The integrity of the technician is all we can depend on. And integrity in this case translates into quality.

Answer 4: Yes.

Answer 5: Yes.

2. Are the ranking forms appropriate for appraising performance or can you suggest a better approach?

Answer 1: Yes.

Answer 2: They may be useful, but are, by themselves, shallow. In many jobs the technician either can do the job, or he can't with no gradations in quality. Quality in trouble-shooting may not necessarily go with quality of repair. In many jobs, particularly in avionics, poor quality will probably go undetected. Supervisors will probably be influenced by a strong halo effect.

Answer 3: They are O.K. I can't suggest a better one offhand but I would be interested to know how specific subordinates "turned out."

Answer 4: May be appropriate for research--oversimplified for the many jobs in aircraft maintenance.

Answer 5: MSEP data is currently used for this.

3. Are quantity and quality useful measures of performance?

Answer 1: Definitely important inputs in the total evaluation of the individual.

Answer 2: (I guess I started answering this above.) Useful, but if they are to be the only measures, and the rater knows they are the only measures, they will reflect far more than their titles. They will become APR's.

APPENDIX L (CONT.)

SURVEY ANSWERS

3. Are quantity and quality useful measures of performance?

Answer 3: Yes.

Answer 4: Yes, although I am biased toward quality based on my experience in Q.C. Organization measures of performance are also important.

Answer 5: Yes-measures currently in use.

4. Which do you consider to be more important, quantity or quality?

Answer 1: One can hardly be considered without the other. Everything being equal, I would choose quality over quantity.

Answer 2: Quality standards must be met in any job, regardless of quantity (speed). It is more complicated than that, but "quality" comes first.

Answer 3: Quality.

Answer 4: Quality (see #3).

Answer 5: Quality with the exception of wartime conditions--budget is also a factor.

5. If one is more important than the other, can you indicate how much more important it is?

Answer 1: Quality is more important only to the degree that without quality maintenance the mission would be jeopardized, i.e., safety, aborts, out-of-commission rates, etc.. Regardless of the amount of output, if it's not reliable the quantity would do little for mission accomplishment. Quality is considerably more important.

Answer 2: "Quality" is absolutely overriding in importance, BUT "quality" is not an absolute. For example, if a perfectly reliable "temporary fix" saves a mission no one will fault the loss in quality over a lengthy permanent repair. The same component in the shop for overhaul would not be acceptable with the temporary fix. Another example of the nebulous nature of quality might be corrosion control. The Air Force, particularly MAC, wants factory new paint jobs. The technician who does a by-the-book perfect job of inhibiting corrosion on

APPENDIX L (CONT.)

SURVEY ANSWERS

5. prominent panel may, by technical standards have achieved quality, but by other standards, have done a poor job.

Answer 3: Quality is by far more important in most instances. Quantity, in my estimate, is more important in a very few cases, probably in wartime, when battle outcome may depend on speed. In such cases the decrease in quality can only be tolerated in some areas, and is, or should be, a calculated thing.

Answer 4: Quality is always more important with some possible exceptions during wartime.

Answer 5: No answer.

APPENDIX M

Glossary

AFB. (Air Force Base).

AGE (aerospace ground equipment). All equipment required on the ground to make a system operational in its intended environment.

APR (Airmen Performance Rating). An annual or semiannual appraisal of Airmen.

consistent estimator. An estimator is a consistent estimator of a parameter if, with increasing sample size, the probability that the value of the statistic is very near that of the parameter becomes closer and closer to unity.

efficient estimator. An estimator is a more efficient estimator than another if its standard error is smaller for the same sample size.

estimator. A statistic obtained from a sample to estimate a population parameter. For instance, the sample mean is a particularly good estimator for the population mean.

FMS (Field Maintenance Squadron). FMS is responsible for fabrication, engine and aircraft subsystem repair, and AGE.

heteroscedasticity. The case where regression error variance is not constant over all observations.

histogram. A graphical portrayal of a population frequency distribution.

kurtosis. More or less peaked than a normal distribution.

MMICS (Maintenance Management Information and Control System). A base level computer system designed to improve the effectiveness of maintenance organizations.

MSEP (Maintenance Standardization and Evaluation Program). A quantitative quality control program designed to check individual technical competence and the quality of maintenance through evaluations and inspections.

OMS (Organizational Maintenance Squadron). OMS is responsible for aircraft launching and recovery and inspections.

APPENDIX M (CONT.)

Glossary

non-parametric test. A test which makes no hypothesis about the value of a population parameter.

regression analysis. An analysis which indicates how one variable is related to another. It provides an equation wherein the known value of one variable may be used to estimate the unknown value of the other. It is distinct from correlation analysis, which indicates the degree to which two variables are related.

skewed. A population is skewed when the mean, median, and mode do not coincide and the frequency curve has a "tail" tapering off to one side.

SKT (Skill Knowledge Test). A paper and pencil test administered to specified technician specialties prior to promotion evaluations.

Type I error (α). The probability of erroneously rejecting a hypothesis.

Type II error (β). The probability of erroneously accepting a hypothesis.

unbiased estimator. A statistic that has an expected value equal to the population parameter being estimated.

USAF (United States Air Force).

BIOGRAPHICAL SKETCH

Joel R. Hickman was born in Los Angeles, California, on July 20, 1948. He received his elementary education in the Los Angeles Public Schools and his secondary education in the Santa Monica Public Schools. He attended the University of California at Los Angeles and graduated with a Bachelor of Science degree in structural and materials engineering in 1971. In August, 1972, he was commissioned as a Second Lieutenant in the United States Air Force. Since that time he has served as an instructor and flight examiner navigator and wing flight safety officer in the Military Airlift Command at McChord AFB, Washington. While at McChord AFB he entered the off-campus graduate business school of Southern Illinois University, Edwardsville, and graduated with a Master of Business Administration degree in 1978. In 1978 the Air Force selected him to pursue a Master of Science degree in Industrial and Management Systems Engineering at Arizona State University. Following the completion of the Master's degree, Captain Hickman will serve with the Civil Engineering Squadron, K.I. Sawyer AFB, Michigan. He is a member of Beta Gamma Sigma, the business and management honor society, and Alpha Pi Mu, the industrial engineering honor society. He is married and the father of a son and daughter.